

Virulence Analysis Tool (VAT)

User Manual

Antje Herrmann

Christian-Albrechts-Univ. Kiel,
Institute of Crop Science and Plant
Breeding

Amos Dinoor

The Hebrew Univ. of Jerusalem
Faculty of Agricultural, Food and
Environmental Quality Sciences

Gabriel A. Schachtel

Justus-Liebig Univ. Giessen
Biometrie & Pop.-Genetik, FB 09

Evsey Kosman

Tel Aviv University,
Institute for Cereal Crops Improvement

March 2009

This project was supported by the *German-Israeli Foundation (GIF)*.

§1. INTRODUCTION	1
1.1 General information about VAT	1
1.2 Using VAT?.....	2
Getting started	2
Basic tools and displays.....	2
1.3 VAT Main Window.....	3
Type of Analysis sector in the VAT Main Window.....	4
Applications sector in the VAT Main Window.....	4
Part I: Virulence Analysis applications	5
§2. DATA ENTRY	5
2.1 Create Differential Set.....	6
2.2 Types of Data	8
2.3 Original Data Sheet	12
2.4 Enter New Data	22
2.5 Open Existing File.....	25
§3. RESAMPLING AND CODING	28
3.1 The Files Selection Sheet	29
Conversion by Transfer	30
3.2 The Coding function.....	32
The Original Input Sheet	32
The Binary Representation Sheet.....	33
The Coded Representation Sheet.....	34
The Comments Sheet.....	36
3.3 Resampling Function.....	36
§4. DESCRIPTIVE STATISTICS	38
4.1 The Files Selection Sheet	38
4.2 Within-population analysis.....	40
The Phenotype Characterization Sheet.....	41
The Phenotype Frequencies Table.....	41
The Three Implemented Race Codes	42
The Virulence Complexities Table.....	42
The Comparison of Individuals Sheet	42
The Virulence Frequency Sheet	43
The Comparison of Differentials Sheet.....	44
The Diversity Parameters Sheet	45
The Comments Sheet.....	45

4.3 Between-Populations Analysis	45
The Pairwise Common Phenotype Sheet	46
The Comparison of Phenotypes Sheet.....	47
The Overall Common Phenotypes Sheet.....	48
The Virulence Frequency Sheet	49
The Between-Sample Distance Sheet.....	50
The Comments Sheet.....	50
§5. INFERENCE STATISTICS	51
5.1 The Files Selection Sheet	51
5.2 Within-Population Analysis	53
The Phenotype Characterization Sheet.....	53
The Phenotype Frequency Table	55
The Virulence Complexities Table.....	55
The Virulence Frequency Sheet	56
The Comparison of Differentials Sheet.....	56
The Diversity Parameters Sheet	57
The Comments Sheet.....	58
5.3 Between-populations analysis	58
The Virulence Frequency Sheet	58
The Between-Population Distances Sheet.....	59
The Comments Sheet.....	60
Part II: Resistance Analysis applications	61
§6. RESAMPLING AND CODING FOR RESISTANCE DATA.....	61
§7. APPENDIX	64
§8. REFERENCES	76

§1. Introduction

The analysis of plant pathogen populations is commonly based on experimental data which are organized in large two-way tables. The *Virulence Analysis Tool* (VAT) is user friendly software for processing such kind of data. VAT aims at supporting a comprehensive, effective and logically consistent evaluation and presentation of virulence data of pathogen populations and of resistance data of host populations. The package can also be applied to molecular marker data. VAT offers the following features:

- (1) Tools to facilitate basic routine steps such as data entry and transformation, dichotomization, identification of phenotypes. A tool to translate phenotype (race) names from one nomenclature to another, e. g. from binary/octal (Gilmour code) to binary/hexadecimal (Roelfs/McVey code) is implemented to make results of different researchers compatible.
- (2) **Descriptive tools** for characterization of isolate and host samples (e.g. by distribution of phenotypes, virulence/resistance frequencies and complexities, associations, diversities, distances etc), displaying the results in a clear and organized fashion by histograms (under development), frequency tables and indices.
- (3) **Inference-statistical** procedures that estimates various diversity and distance indices and other parameters for sexually and asexually reproducing populations. These estimates are obtained by resampling methods allowing further statistical evaluation (e.g. significance tests and confidence intervals).
- (4) **Sample size** recommendations for reliable estimation in specific experimental situations are provided.

VAT output is **compatible** to all major statistical analysis tools and suitable for direct input into MS Excel and most other commonly used packages (SAS, NTSYS, SPSS etc) facilitating additional analyses (clustering, dendrograms, PCA etc).

1.1 General information about VAT

Input files. The program accepts only the standard text files (**inputfilename.txt** with extension **txt**) or the system oriented files (**filename.vat** with extension **vat** for virulence data and **filename.rat** with extension **rat** for resistance data) which are

created by the VAT program itself. The txt-files will only be accepted in the Data Entry section. The system vat-files and rat-files can be read at any phase or section of the program for virulence and resistance data, respectively, and data analysis is possible only with vat-files and rat-files.

Output files. The basic data output of the program is a system vat(rat)-file. User should create the **filename.vat** (**filename.rat**) file in order to allow data processing. This file contains original data, results of data transformation and resampling, and some other information on original data. However, the most important results of calculations obtained in the various sections (e.g. **Descriptive Statistics**, **Inferential Statistics**, and **Miscellaneous**) are **not saved** in the corresponding vat(rat)-file. The major results of the calculations, as well as original and transformed (encoded) data, can be exported to **Excel** at every step in the program for further study. Data can also be saved as a txt-file.

Another option **Save to File** is available within the sections. By clicking this option you can open a new file and save the results of calculations as a standard text file (**outputfilename.txt**, i. e. with extension **txt**). However, the readability of the **outputfilename.txt** might be poor and need some additional formatting by the user. For example if one of the analyses did not develop results, then a headline without any corresponding results may appear in the text file. Moreover, sometimes columns and titles will not be aligned. Therefore, this option should be used mainly for backup purposes.

1.2 Using VAT?

Getting started

Startup. In order to run the VAT, go into a folder where VAT software (VATsoftware folder) was installed (downloaded) and look for the executable file VAT.exe according to the following route: VATsoftware\VAT\bin\VAT.exe. Click VAT.exe, and the VAT Main Window should appear. It is recommended to create a shortcut of VAT.exe, and include the corresponding icon on the desktop to start the VAT.

Basic tools and displays

The VAT program allows the analysis of pathogenic data in both directions, namely **Virulence Analysis** and **Resistance Analysis**. For each type of analysis there are five

applications: (1) **Data entry**, (2) **Resampling and Coding**, (3) **Descriptive Statistics**, (4) **Inferential Statistics** and (5) **Miscellaneous** (under development), which will be described in more detail in the following chapters.

VAT displays data, comments and results of calculations mainly in grids, tables and matrices. In the **Data Entry** section you can enter and edit your data and comments into grids and tables. In all other sections the contents of grids, tables and matrices cannot be modified by the user.

Format of grid cells (height of rows and width of columns) can be changed by dragging the corresponding borders. Most windows or sheets in VAT have an **Excel** icon which allows a user to copy data and results to an Excel worksheet.

The **General Management Bar** can be found above most VAT windows. It contains the options **File**, **Change Application** and **Help** (under development). Under **File** there is the **Exit** option to leave the program and terminate current VAT session. With the **Change Application** option one can switch to other VAT applications or return to the **VAT Main Window**. The **Help** button provides information how to work with VAT.

1.3 VAT Main Window

The **VAT Main Window** is divided into two sectors: **Type of Analysis** (at the top) and **Applications** (the main square) (Fig. 1.1). It is possible to activate any of the different applications in the **Applications** sector combined with each of the two types of analysis.

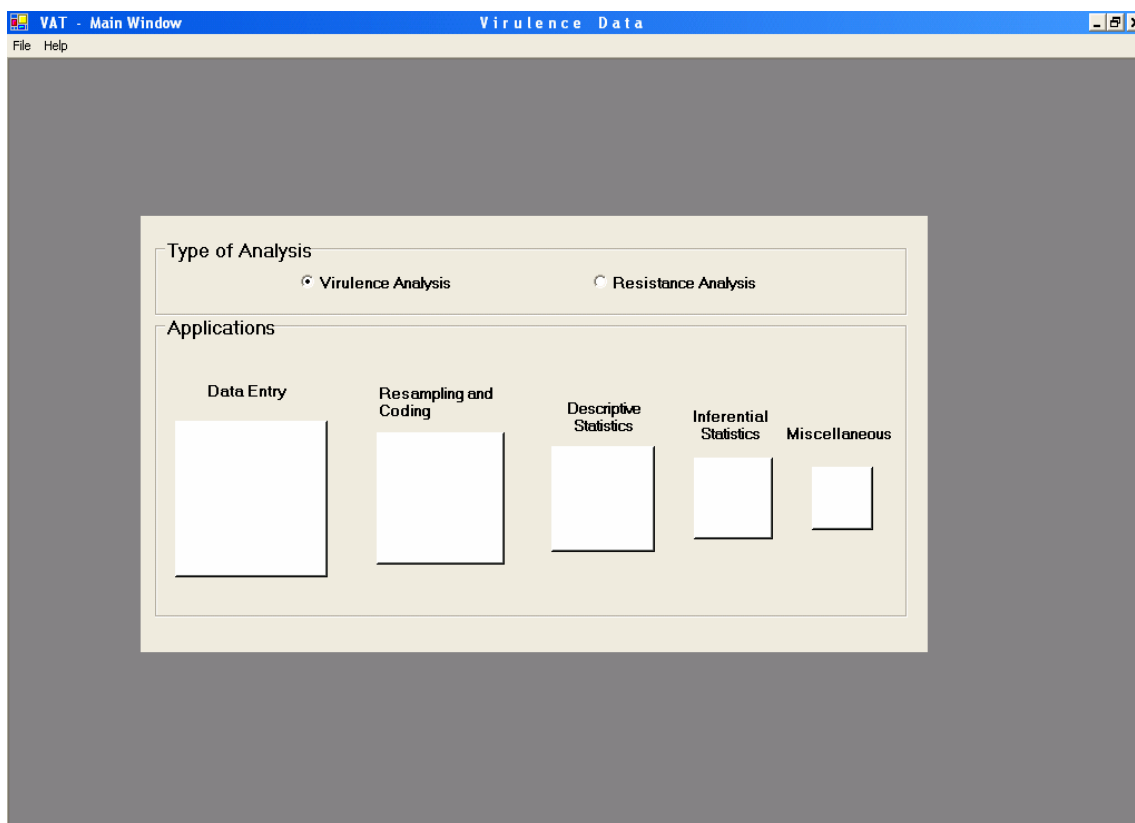


Figure 1.1: VAT Main Window. At the top is the **Type of Analysis** sector where the **Virulence Analysis** option is chosen. In the **Applications** sector there are the five squares representing available applications.

Type of Analysis sector in the VAT Main Window

VAT provides two types of analysis: **Virulence Analysis** and **Resistance Analysis**. These options are available at the top sector (Type of Analysis) and can be chosen by clicking the appropriate circle.

Virulence Analysis - Virulence patterns of isolates with respect to a given differential set *of hosts* will be analyzed. One can characterize and compare individual isolates and isolate populations tested for virulence on a given differential set.

Resistance Analysis - Resistance patterns of host plants with respect to a given differential set *of isolates* are analyzed. One can characterize and compare individual hosts and host populations tested for resistance on given differential set.

Applications sector in the VAT Main Window

There are five different applications: (1) **Data entry**, (2) **Resampling and Coding**, (3) **Descriptive Statistics**, (4) **Inferential Statistics** and (5) **Miscellaneous** (under development), represented by differently sized squares. Each application is activated

by clicking on the desired square. In principle, these applications are arranged in the logical succession of steps of analysis, starting from the biggest square (**Data entry**) in the left.

Part I: Virulence Analysis applications

§2. Data Entry

This VAT application allows user to perform three operations, namely

1. Create new and/or modify existing differential sets. Differential sets can be saved in the program and used for further data entry.
2. Enter new data and save it for further analysis. New data appear in a grid (table) where rows represent virulence patterns (original or encoded) of isolates for a given differential set. The data type could be either: Regular, Binary, Octal, Hexadecimal, or Binary-Decimal.
3. Open an existing file and modify it for further analysis. Two file types are allowed: either the text txt-file or the system vat-file.

By choosing the **Data Entry** square in the **VAT Main Window**, the **Data Entry** window will open.

The **Data Entry** window (Fig. 2.1) displays the three options: **Create Differential Set**, **Enter New Data**, and **Open Existing File**, each represented by a square. Above these windows is the **General Management Bar** with buttons **File**, **Change Application** and **Help**. Under **File** there is the **Exit** option which allows terminating the program. **Change Application** allows a user to switch to other **VAT Application** or to return to the **VAT Main Window**.

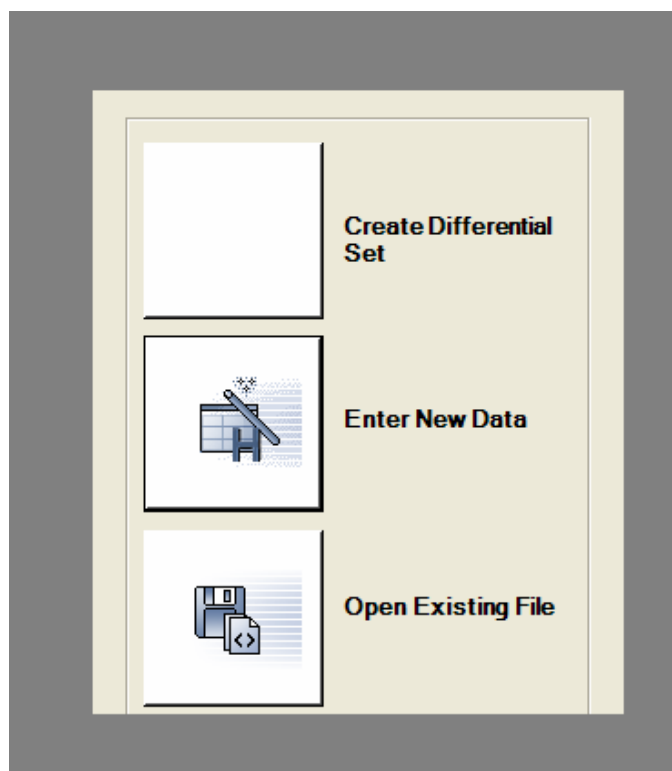


Figure 2.1: **Data Entry** window. Three squares represent three different options namely, Create Differential Set, Enter New Data, and Open Existing File.

2.1 Create Differential Set

This option allows you to customize your own **Differential Set** or modify an existing **Differential Set**. Notice that term **Differential Set** means an ordered set of differentials. If the same differentials are placed in different order, then the corresponding differential sets are different. The **Differential Set** window is shown in Fig. 2.2.

Differential Set

DiffSet 6.xml **UPDATE existing set**

Number of differentials (columns): 7 **Add Ins Del**

Name: DiffSet 6 **Country:** Country6
Pathogen: pathogen6 **Reference:** Reference6
Host: host6 **Author:**

Column names

#	Name of Differential	Comment
1	Diff1	comment1
2	Diff2	comment2
3	Diff3	comment3
4	Diff4	comment4
5	Diff5	comment5
6	Diff6	comment6
7	Diff7	this is where you enter commoents

Save **Back to Data Entry** **Help**

Figure 2.2: Differential Set window. Differential set "DiffSet 6" was chosen (DiffSet 6.xml is name of the corresponding system file which contains "DiffSet 6"). This differential set has 7 columns named Diff1, Diff2, etc. Each differential can be commented as comment1, comment2, etc. The marker here is set to Diff7.

The field in the top center allows either to browse and select an already existing differential set or to choose the option **New** and to enter a new differential set.

New Differential Set. Once the **New** option is chosen the field **Number of differentials** (columns) will be available. Enter the number of differentials (hosts, isogenic lines etc) in the new set and click **Go**. A grid in the **Columns Name** box will be created, where number of rows equals the number of differentials (number of columns in tables of regular and binary data). Attributes of the differential set (**Name**, **Pathogen**, **Host**, **Country**, **Reference** and **Author**) must/can be entered in the appropriate slots. Differential names and comments can be inserted in the corresponding grid cells in the **Column name** box.

Once the differential set characterization is completed, click the **Save** button at the bottom in order to save your differential set under the given name. Notice that a

user-defined **Name** of the differential set is mandatory in order to save it. Otherwise once you attempt to save, an error message will appear. After saving the differential set will be available for use in the **Enter New Data** section (see §2.4).

Existing Differential Set. Selection of an existing differential set in the top box of the **Differential Set** window will automatically display all available information about the set (attributes, names of differentials and comments). In addition to editing the existing information, the possibility to modify the differential set is facilitated by the **Insert**, **Add** and **Del** buttons which will appear next to the field **Number of differentials** (columns).

Insert. Mark any row in the list of differentials from the **Columns name** box. An arrow will appear next to the marked row. By clicking the **Insert** button, a new row for entering name and comment for the new differential will be inserted above the marked one.

Add. By clicking **Add**, a new row for entering name and comment for the new differential will be added at the bottom in the **Columns name** box.

Del. By clicking **Del**, the marked differential will be deleted from the list in the **Columns name** box.

The entries of **Name**, **Pathogen**, **Host**, **Country**, **Reference** and **Author** can be modified by clicking into the appropriate box. The name and comments of each member of the differential set can be modified by clicking the appropriate cell at the **Column name** box. Once editing and modification of the differential set is completed, click the **Save** button at the bottom. Now the differential set is saved by VAT under the chosen **Name** (original or modified), and is available in the **Enter New Data** section (see §2.4).

Clicking the **Excel** icon at the top right will transfer the current data to an Excel file and open this file.

The **Back to Data Entry** button is for return from the **Differential Set** window to the **Data Entry** window.

2.2 Types of Data

The following five types of data are admissible.

Regular data. If number of individuals and differentials in data table equal k and n , respectively, then **Regular data** are determined as $k \times n$ table of nonnegative real

numbers with k rows and n columns. Each row represents a reaction pattern (vector-pattern) of an individual on the given set of differentials according to the assessment scale underlying the data. Nonnegative real numbers within the range of assessment scale are the only valid entries in the case of **Regular data**.

Binary data. If number of individuals and differentials in data table equal k and n , respectively, then **Binary data** are determined as $k \times n$ table of 0s and 1s with k rows and n columns. Each row represents a reaction pattern (binary vector-pattern) of an individual on the given set of differentials, where 0 or 1 at i -th position corresponds to avirulence (negative) or virulence (positive) reaction, respectively, of the individual on i -th differential. Two symbols 0 and 1 are the only valid entries in the case of **Binary data**.

Octal data. If number n of differentials in a **Differential Set** is a multiple of three (i.e. $n = 3l$), then a binary reaction pattern of an individual on the given set of differentials can be represented by the **Octal Code** as follows. The ordered set of differentials is divided in l separate groups of three consecutive differentials each. Then binary patterns of individuals are represented by the ordered sets of l $\{0,1\}$ -triplets (triplet means three ordered symbols) according to the division of differentials into groups. Now each triplet can be encoded by a single digit from 0 to 7 according to the following rule:

Triplet	Octal Code	Rule
000	0	$0 = 0 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0$
001	1	$1 = 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$
010	2	$2 = 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0$
011	3	$3 = 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$
100	4	$4 = 1 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0$
101	5	$5 = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$
110	6	$6 = 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0$
111	7	$7 = 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$

This is one-to-one correspondence between the eight digits and all possible $\{0,1\}$ -triplets. Substituting one of the digits from 0 to 7 for the corresponding triplet in the binary pattern of individual, the individual's **Octal Code** of length l is obtained. Obviously, any **Octal Code** can be transformed to the binary vector according to the

same rule. For example, the following pairs of binary vectors of length 15 and octal codes of length 5 are results of the corresponding transformations:

$$35164 \leftrightarrow 011,101,001,110,100 \leftrightarrow 011101001110100;$$

$$13704 \leftrightarrow 001,011,111,000,100 \leftrightarrow 001011111000100.$$

If number of individuals in data table equals k , then **Octal data** are determined as a table of k rows and one column of octal codes of a fixed length. Digits $\{0,1,2,3,4,5,6,7\}$ are the only valid symbols in the case of **Octal data**, and all entries should have the same number of symbols (fixed **Code length**).

Hexadecimal data. If number n of differentials in a **Differential Set** is a multiple of four (i.e. $n = 4l$), then a binary reaction pattern of an individual on the given set of differentials can be represented by the **Hexadecimal Code** as follows. The ordered set of differentials is divided in l separate groups of four consecutive differentials each. Then binary patterns of individuals are represented by the ordered sets of l $\{0,1\}$ -four-tuples (four-tuple means four ordered symbols) according to the division of differentials into groups. Now each four-tuple can be encoded by a single letter from the set of the first sixteen consonants of English alphabet $\{B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T\}$ according to the following rule:

Four-tuple	Hexadecimal Code	Four-tuple	Hexadecimal Code
0000	B	1000	L
0001	C	1001	M
0010	D	1010	N
0011	F	1011	P
0100	G	1100	Q
0101	H	1101	R
0110	J	1110	S
0111	K	1111	T

This is one-to-one correspondence between the sixteen letters and all possible $\{0,1\}$ -four-tuples. Substituting one of the letters for the corresponding four-tuple in the binary pattern of individual, the individual's **Hexadecimal Code** of length l is obtained. Obviously, any **Hexadecimal Code** can be transformed to the binary vector according to the same rule. For example, the following pairs of binary vectors of length 24 and hexadecimal codes of length 6 are results of the corresponding transformations:

BNJHGR \leftrightarrow 0000,1010,0110,0101,0100,1100 \leftrightarrow 000010100110010101001100;

DDTGLK \leftrightarrow 0010,0010,1111,0100,1000,0111 \leftrightarrow 001000101111010010000111.

If number of individuals in data table equals k , then **Hexadecimal data** are determined as a table of k rows and one column of hexadecimal codes of a fixed length. The first sixteen consonants of English alphabet {B, C, D, F, G, H, J, K, L, M, N, P, R, S, Q, T} are the only valid symbols in the case of **Hexadecimal data**, and all entries should have the same number of symbols (fixed **Code length**).

Binary-Decimal data. If a **Differential Set** consists of n differentials, then a binary reaction pattern of an individual on the given set of differentials has the following general form: $\mathbf{v} = (v_1 v_2 v_3 \dots v_{n-1} v_n)$, where $v_i = 0$ or $v_i = 1$ for all $i = 1, 2, \dots, n$. One-to-one correspondence of such patterns to integers between 0 and $2^n - 1$ is naturally established by matching the integer $I(\mathbf{v}) = v_1 \cdot 2^{n-1} + v_2 \cdot 2^{n-2} + v_3 \cdot 2^{n-3} + \dots + v_{n-1} \cdot 2^1 + v_n \cdot 2^0$ to the binary pattern \mathbf{v} . This integer $I(\mathbf{v})$ is called the **Binary-Decimal Code** of \mathbf{v} . Obviously, any **Binary-Decimal Code** can be transformed to the single binary vector of length n according to the same rule. For example, in the case of six differentials ($n = 6$) the following binary vectors are represented by the corresponding binary-decimal Codes:

$$010101 \leftrightarrow 21 = 0 \cdot 2^5 + 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0;$$

$$101101 \leftrightarrow 45 = 1 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0;$$

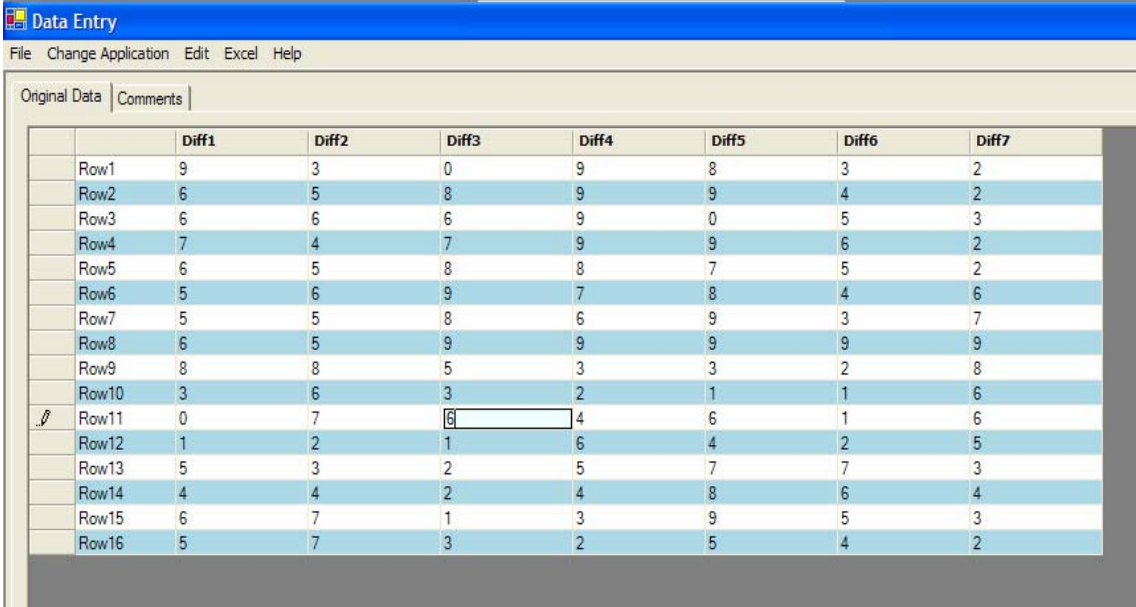
$$000011 \leftrightarrow 3 = 0 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0.$$

If number of individuals in data table equals k , then **Binary-Decimal data** are determined as a table of k rows and one column of binary-decimal codes.

2.3 Original Data Sheet

Original Data Sheet is the main instrument for entering, editing and modifying data and comments, as well as for displaying data, data parameters (type of data, size of data table etc), and name and location of data file. Data always appear in the grid, while information about the data and data file is placed below the grid. Note that only valid values for declared type of data can appear or can be entered in the grid. Otherwise, in the case of invalid values, error messages will pop up. For example, for regular data with specified range of assessment scale between minimum and maximum values (let say 1 and 10, respectively), neither numbers smaller than 1 or greater than 10, nor letters, symbols etc. cannot be entered.

Pencil Mode. Once entering a cell (clicking in the grid cell) and typing there, the regular row marker (a triangle pointing to the right) will turn into a pencil with three dots (see Fig. 2.3 row 11). This mode is called the **Pencil Mode**. In this mode the program enables entering data. When typing within the cell is completed, quit this cell (click anywhere outside the cell) in order to return to the **Regular Mode**. Note that in the **Pencil Mode** some functions may be incorrectly executed. Therefore, it is always recommended to leave the **Pencil Mode** before continuing processing or saving.



	Diff1	Diff2	Diff3	Diff4	Diff5	Diff6	Diff7
Row1	9	3	0	9	8	3	2
Row2	6	5	8	9	9	4	2
Row3	6	6	6	9	0	5	3
Row4	7	4	7	9	9	6	2
Row5	6	5	8	8	7	5	2
Row6	5	6	9	7	8	4	6
Row7	5	5	8	6	9	3	7
Row8	6	5	9	9	9	9	9
Row9	8	8	5	3	3	2	8
Row10	3	6	3	2	1	1	6
Pencil Row11	0	7	6	4	6	1	6
Row12	1	2	1	6	4	2	5
Row13	5	3	2	5	7	7	3
Row14	4	4	2	4	8	6	4
Row15	6	7	1	3	9	5	3
Row16	5	7	3	2	5	4	2

Figure 2.3: Original Data Sheet. Regular data is filled in here. In cell (Row11,Diff3) the number 6 has just been entered, therefore the program is in the **Pencil Mode** (marked next to row 11).

The **General management bar** at the top of the **Data Entry** window includes the following five options: **File**, **Change Application**, **Edit**, **Excel**, and **Help**.

File. Under this label there are three alternatives namely: **Save**, **Save as** and **Exit**.

Save. This option is for saving data as a system vat-file (filename.vat; see §1.1). Once the **Save** option is chosen, to different modes of action are realized. If the original data were imported from already existing vat-file, then a message window will appear asking whether to overwrite the existing file. If new data are to be saved in a new vat-file, then the standard Microsoft **Save as** window will appear (Fig. 2.4) to choose the file name and path. Note once again that only vat-files can be saved here. In all **VAT Applications** (except **Data Entry**) only vat-files can serve as input and be processed.

To save data as a text file (filename.txt), choose the **Save As** option.

Save As. This option is for saving data as a system vat-file (filename.vat; see §1.1) under a new name or as a text file (filename.txt). By choosing this option a system special **Save as** window will appear (Fig. 2.5). It is mandatory to select **Type of File** (system **vat-file** or text **txt-file**), to choose the file name and path clicking the **Browse** button, and to select a **Column delimiter** in the case of text **txt-file** with **Binary** or **Regular** data (see §2.2). Once finished, click **OK** to save the data under the file in the File Name box.

Type of File selection.

System vat-file. Selection of "System VAT" provides saving data in a special system oriented file format under any name with a fixed extension vat (filename.vat). The vat-files are readable and can be used in all **VAT Applications**. Click the **Browse** button to choose the file name and path. The standard Microsoft **Save as** window will appear (Fig. 2.4) in order to browse through for saving the data in an appropriate folder under any name but with only possible extension vat (filename.vat). Only folders and/or the system vat-files will be shown in the **Save as** window (Fig. 2.4).

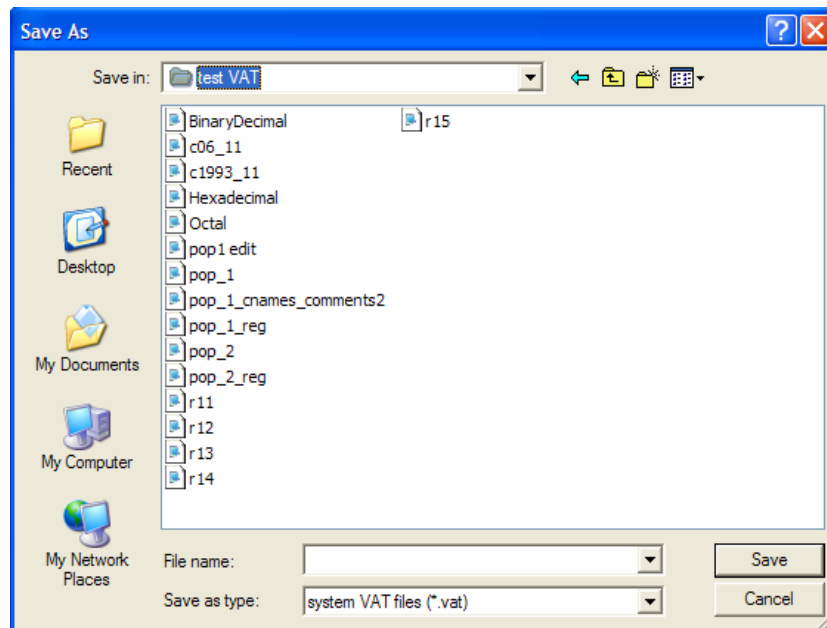


Figure 2.4: Save as window (standard). The folder's path ends in the folder test VAT. File name has to be entered.

Text txt-file. Selection of "Text File" provides saving data in the standard text file format under any name with the standard extension txt (filename.txt). This option is mainly for backup or importing/exporting data. Note that txt-files are not readable and cannot be used in all available **VAT Applications** (except **Data Entry**). Click the **Browse** button to choose the file name and path. The standard Microsoft **Save as** window will appear (Fig. 2.5a) in order to browse through for saving the data in an appropriate folder under any name but with only possible extension txt (filename.txt). Only folders and/or the text txt-files will be shown in the **Save as** window (Fig. 2.4). Determining a **Column delimiter** in the case of **Binary** or **Regular** data (see §2.2) is mandatory.

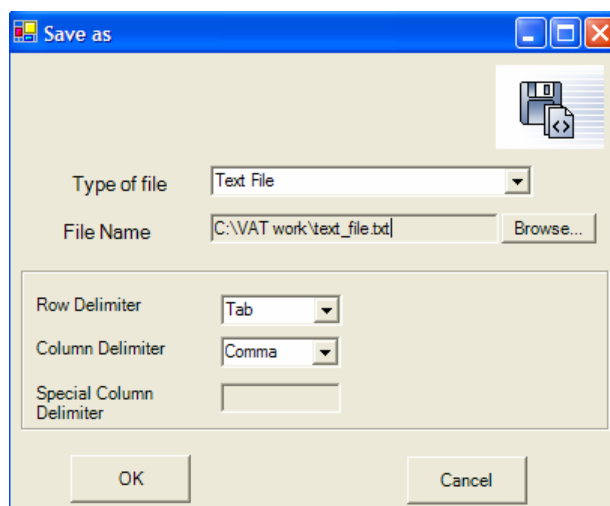


Figure 2.5: Save as window (system). Text File is chosen and the file path is C:\VAT work\text_file.txt. the **Row delimiter** is Tab and the Column Delimiter is a Comma.

Change Application. This menu option provides a switch among the five VAT Applications (**Data entry, Resampling and Coding, Descriptive Statistics, Inferential Statistics, Miscellaneous**) and the **Main Window**.

Edit. This menu option provides various tools (functions) for editing, modifying and validation of original data: **Transposition, Validate, Reset, Replace, 0-1 Transformation, Add, Insert, Delete, Rename, and Missing Data.**

Transposition. This function is for transposition of a data table in the grid of **Original Data Sheet** (rows become columns and vice versa).

Validate. This tool allows checking validity of data in the grid of **Original Data Sheet**. It is mainly used in order to find empty cells in large arrays of data and to detect incomplete codes in the case of **Octal** and **Hexadecimal data**. Note that all other cases of invalid data are automatically protected on the stage of entering new data or importing data from existing files. Missing data (empty cells) are allowed (valid) in the **Data Entry** application, but processing files with missing data (incomplete data) in all other **VAT Applications** is impossible. The **Validate** function has two options: **List invalid cells** and **Find next invalid cell**.

List invalid cells. This tool provides information about all invalid cells. **Validation** window will be opened to display coordinates (row number, row name, column name, and column number) of cells with missing data or incomplete codes (see Figs. 2.6a and 2.6b with one empty cell (Row12,Diff4)). To save the displayed **List of Invalid Cells** click the **Save** button and the standard **Save As** window will appear (Fig. 2.4).

File Change Application Edit Excel Help								
Original Data		Comments						
		Diff1	Diff2	Diff3	Diff4	Diff5	Diff6	Diff7
▶	Row1	9	3	0	9	8	3	2
	Row2	6	5	8	9	9	4	2
	Row3	6	6	6	9	0	5	3
	Row4	7	4	7	9	9	6	2
	Row5	6	5	8	8	7	5	2
	Row6	5	6	9	7	8	4	6
	Row7	5	5	8	6	9	3	7
	Row8	6	5	9	9	9	9	9
	Row9	8	8	5	3	3	2	8
	Row10	3	6	3	2	1	1	6
	Row11	0	7	6	4	6	1	6
	Row12	1	2	1		4	2	5
	Row13	5	3	2	5	7	7	3
	Row14	4	4	2	4	8	6	4
	Row15	6	7	1	3	9	5	3
	Row16	5	7	3	2	5	4	2

Figure 2.6a: Regular data with one empty cell (Row12,Diff4).

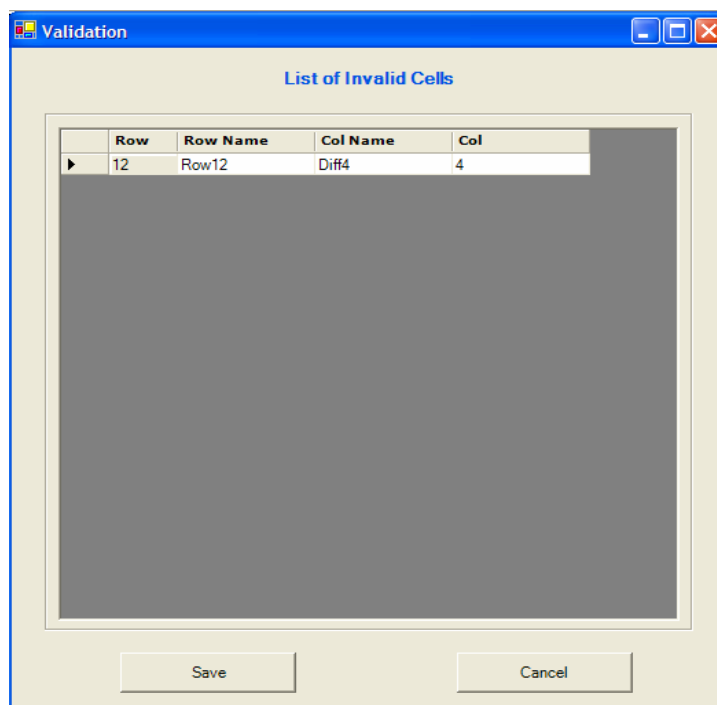


Figure 2.6b: Validation window corresponds to the data from Fig. 2.7a with one empty cell (Row12,Diff4). **List of Invalid Cells** contains information about the cell with missing data (row number, row name, column name, and column number).

Find next invalid cell. By clicking this function or by pressing **F3** button one can scroll through the invalid cells.

Reset. This option allows clearing all data in the grid of **Original Data Sheet**. All cells will become empty.

Replace. This function is available only for **Regular data**. It enables to change all entries with a specific value for another valid value. The **Replace** window will be

opened (Fig. 2.7). Enter the value which is destined to be replaced into the "Find" field, and enter the new value to replace the old one into the "Replace with" field.

	Diff1	Diff2	Diff3	Diff4	Diff5	Diff6	Diff7
Row1	9	3	0	9	8	3	2
Row2	6	5	8	9	9	4	2
Row3	6	6	6	9	0	5	3
Row4	7	4	7	9	9	6	2
Row5	6	5	8	8	7	5	2
Row6	5	6	9	7	8	4	6
Row7	5	5	8	6	9	3	7
Row8	6	5	9	9	9	9	9
Row9	8	8	5	3	3	2	8
Row10	3	6	3	2	1	1	6
Row11	0	7	6	4	6	1	6
Row12	1	2	1	4	4	2	5
Row13	5	3	2	5	7	7	3
Row14	4	4	2	4	8	6	4
Row15	6	7	1	3	9	5	3
Row16	5	7	3	2	5	4	2

a

b

	Diff1	Diff2	Diff3	Diff4	Diff5	Diff6	Diff7
Row1	3	3	0	3	8	3	2
Row2	6	5	8	3	3	4	2
Row3	6	6	6	3	0	5	3
Row4	7	4	7	3	3	6	2
Row5	6	5	8	8	7	5	2
Row6	5	6	3	7	8	4	6
Row7	5	5	8	6	3	3	7
Row8	6	5	3	3	3	3	3
Row9	8	8	5	3	3	2	8
Row10	3	6	3	2	1	1	6
Row11	0	7	6	4	6	1	6
Row12	1	2	1	4	4	2	5
Row13	5	3	2	5	7	7	3
Row14	4	4	2	4	8	6	4
Row15	6	7	1	3	3	5	3
Row16	5	7	3	2	5	4	2

c

Figure 2.7: Original data (a), **Replace** window (b), and the modified data after replacement (c). In the **Replace** window (b) the "Find" and "Replace with" fields contain 9 and 3, respectively. The modified data (c) have all the cells with the value 9 replaced by the value 3 (for example, cell (Row2, Diff4)).

0-1 Transformation. This function is available only for **Binary data**. It switches all 0s to 1 and all 1s to 0.

Add. There are two choices under **Add**: add **Row** or add **Column**. Each automatically adds a new row or column after the last one, respectively, in the bottom or to the right of the current grid.

Insert. There are two choices under **Insert**: insert **Row** and insert **Column**. In order to insert new column (row) to the left (above) a target column (row), click in any cell of the target column (row) and choose the **Column (Row)** option under **Insert**. A new column (row) will be inserted to the left (above) the marked column (row).

Delete. There are two choices under **Delete**: delete **Row** and delete **Column**. In order to delete a target column (row), click in any cell of the column (row) and choose the **Column (Row)** option under **Delete**. The marked column (row) will be deleted.

Rename. There are two choices under **Rename**: rename **Row** and rename **Column**. Once choosing the **Column (Row)** option, the **Rename** window will be opened displaying the List of Columns (Rows) of the original data and their names in the current order (Fig. 2.8). The name of each column (row) can be modified. Click **OK** in order to fix the changes.

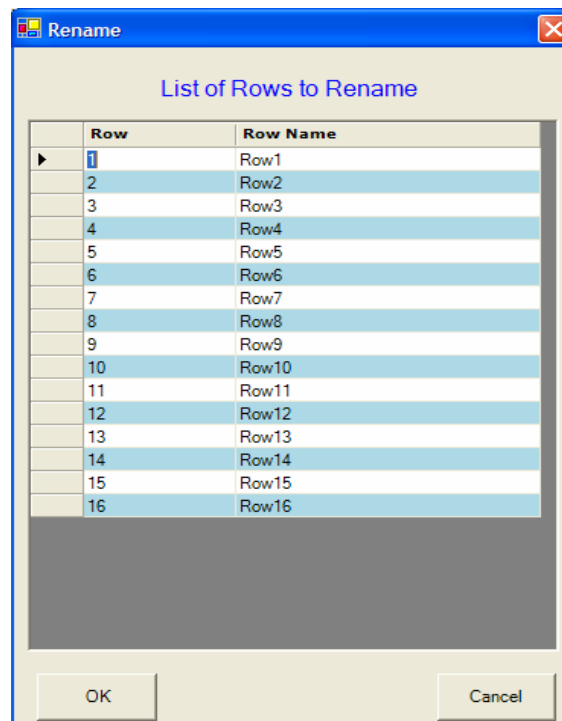


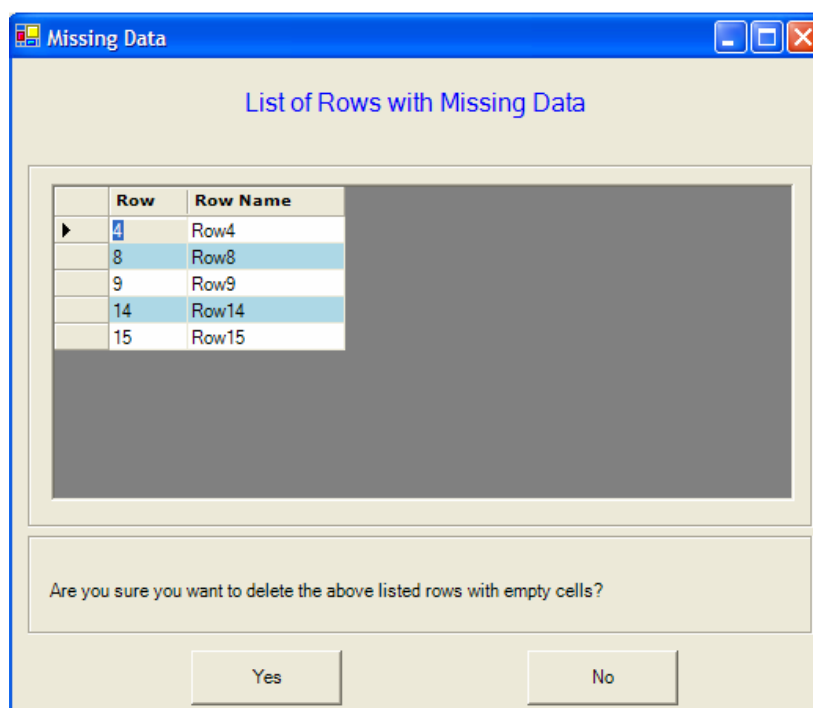
Figure 2.8: Rename window. Names of the 16 rows of original data (Fig. 2.7a) are displayed. The current row names Row1, Row2, Row3 etc can be changed.

Missing Data. This tool is mainly aimed at modification of original data in order to make them acceptable to the computational **VAT Applications: Resampling and Coding, Descriptive Statistics, Inferential Statistics, and Miscellaneous**. This can be achieved by means the following four functions: **Delete all rows with empty cells**, **Delete all columns with empty cells**, **Fill all empty cells with**, and **Delete all rows with incomplete codes** (for encoded data).

Delete all rows with empty cells. This function enables to detect all rows with empty cells (at least one missing entry in a row) and to delete them. By choosing this option the **Missing Data** window will be opened to display all the rows with at least one empty cell (Fig. 2.9b for the corresponding data in Fig. 2.9a). Click **Yes** in order to proceed and to get the modified data (Fig. 2.9c).

	Diff1	Diff2	Diff3	Diff4	Diff5	Diff6	Diff7
Row1	3	4	5	3	2	8	9
Row2	7	6	4	8	2	9	8
Row3	7	8	7	6	7	4	7
Row4	6	8		5	6	3	7
Row5	5	7	6	4	5	6	6
Row6	4	6	4	3	4	8	5
Row7	1	5	3	2	3	9	4
Row8	1		2	3	2	4	1
Row9	2		2	4	2	3	3
Row10	0	5	3	6	2	2	2
Row11	9	4	4	8	2	3	9
Row12	0	3	6	0	2	4	0
Row13	9	4	7	0	4	2	9
Row14	8	5	8		5	3	8
Row15	6	6	9		6	4	7
Row16	5	7	8	9	7	6	6

a



b

		Diff1	Diff2	Diff3	Diff4	Diff5	Diff6	Diff7
	Row1	3	4	5	3	2	8	9
	Row2	7	6	4	8	2	9	8
	Row3	7	8	7	6	7	4	7
	Row5	5	7	6	4	5	6	6
	Row6	4	6	4	3	4	8	5
	Row7	1	5	3	2	3	9	4
	Row10	0	5	3	6	2	2	2
	Row11	9	4	4	8	2	3	9
	Row12	0	3	6	0	2	4	0
	Row13	9	4	7	0	4	2	9
►	Row16	5	7	8	9	7	6	6

c

Figure 2.9: Delete all rows with empty cells. (a) Original data with empty cells; (b) **Missing Data** window displays all rows with empty cells (missing data); (c) Modified data table (five rows were deleted).

Delete all columns with empty cells. This function enables to detect all columns with empty cells (at least one missing entry in a column) and to delete them. By choosing this option the **Missing Data** window will be opened to display all the columns with at least one empty cell. Click **Yes** in order to proceed and to get the modified data (see Fig. 2.9 for rows).

Fill all empty cells with. This function allows to fill in all the empty cells with a specific value. Once chosen, the **Fill Empty Cells** window will be opened (Fig. 2.10). Insert a valid value in the field "Load empty cells with". Once clicking **OK**, all the empty cells in the grid will be filled with this value.

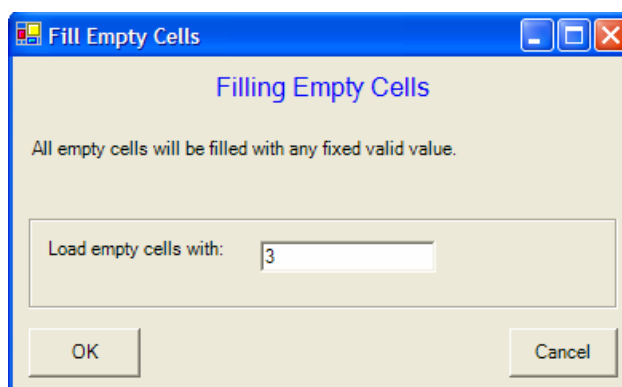


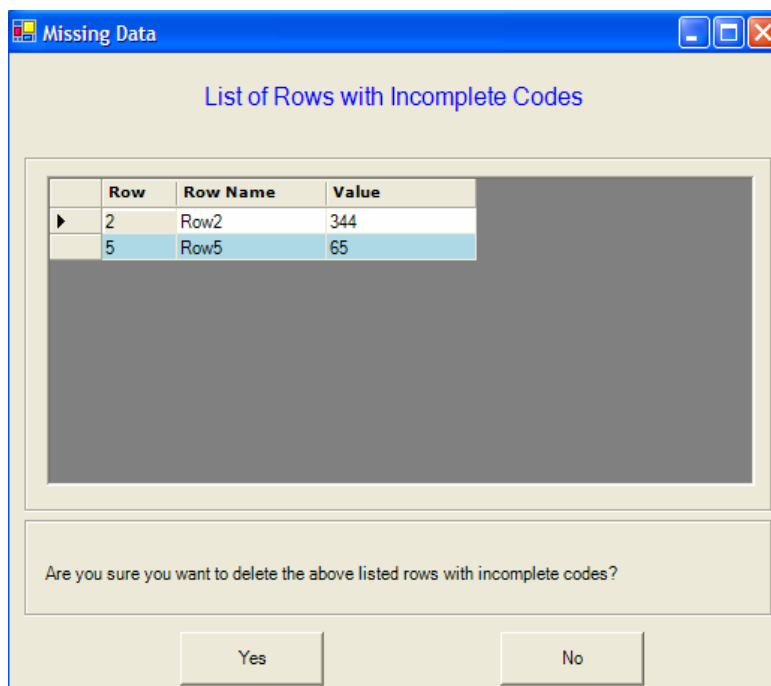
Figure 2.10: Fill Empty Cells window. Value 3 is inserted in the field "Load empty cells with". All empty cells in the grid will be filled with 3.

Delete all rows with incomplete code. This function is available only when using the **Octal data** and **Hexadecimal data**. It allows deleting all rows with incomplete codes. (Incomplete code is the **Octal** or **Hexadecimal Code** with the code length shorter than declared in the **Data Parameters** window (Fig. 2.12) in the case of **Enter New Data**, or shorter than code of maximum length in the case of **Open**

Existing File). Once chosen, the **Missing Data** window will be opened to display all the rows with incomplete code (Fig. 2.11b for the corresponding data in Fig. 2.11a). Click **Yes** in order to proceed and to get the modified data with no incomplete codes (Fig. 2.11c).

		Octal code
	Row1	3625
	Row2	344
	Row3	3674
	Row4	3457
	Row5	65
►	Row6	7653

a



b

		Octal code
	Row1	3625
	Row3	3674
	Row4	3457
►	Row6	7653

c

Figure 2.11: Delete all rows with incomplete code. (a) Original **Octal data** with a specified code length of 4 (i.e. with 12 differentials); (b) **Missing Data** window displays two rows (Row2 and Row5) with detected incomplete codes (344 and 65, respectively); (c) Modified data table (the two rows were deleted).

Excel. This option allows data transfer to the **Excel** worksheet.

2.4 Enter New Data

Enter New Data is the main section in the **Data Entry** application. It provides all necessary tools for generating a new data file, which can be saved and used for subsequent analysis.

By clicking the **Enter New Data** square in the **Data Entry** window (see Fig. 2.1), new window **Data Parameters** (Fig. 2.13) will appear. Three or four parameters must be set, namely (1) **Differential Set**, (2) **Type of Data** and (3) **number of rows (No. rows)**, if an existing differential set is selected, and additional parameter (4) **number of columns (No. cols)**, if none of existing differential sets is selected.

The screenshot shows the 'Data Parameters' dialog box. The 'Differential Set' dropdown is set to 'DiffSet 6.xml'. Under 'Type of Data', the 'Regular' checkbox is selected, with 'Min value' set to 0 and 'Max value' set to 9. The 'Binary' checkbox is unselected. The 'No. cols' field is set to 7. Under 'Code length', both 'Octal' and 'Hexadecimal' checkboxes are unselected, and the 'Code length' spinner is set to 1. Under 'Binary-Decimal', the checkbox is unselected, and the 'Binary vector length' spinner is set to 1. The 'No. rows' field is set to 16. Navigation buttons '< Back' and 'Next >' are at the bottom.

Figure 2.13: **Data Parameters** window. Differential set, DiffSet 6.xml, is selected, and Regular type of data is chosen. The Min value is 0, and the Max value is 9. The number of columns (No. cols.) is 7, and the number of rows (No. rows.) is 16.

Differential Set. Scroll down the field **Differential Set** and select either any differential set from the list of available ones (see §2.1) or None. If any existing **Differential Set** is selected, then the number of differentials in this set will automatically appear in the **No. cols** (number of columns) and the **Binary vector length** fields. If a derivative of the number of differentials is determined for **Octal** or **Hexadecimal data** (see §2.2), it will automatically appear in the **Code length** field.

Type of Data. The next segment of the Data Parameters box requires specifications concerning the **type of data**. One of the following five types should be selected: **Regular**, **Binary**, **Octal**, **Hexadecimal** or **Binary-Decimal** (see §2.2). In general, applications **Descriptive Statistics** (see §4) and **Inferential Statistics** (see §5) will only operate under **Binary data**. Data of other types should be converted to the **Binary** one in the **Resampling and Coding** (see §3).

Regular data. Check the **Regular** box. Then **Min value** and **Max value** fields will be available to specify range of data by entering a minimum value (**Min value**) and a maximum value (**Max value**) according to the assessment scale underlying the data. Notice that the **Min value** and **Max value** are mandatory parameters; they should be arbitrary nonnegative numbers, so that **Min value** is less than **Max value**.

Binary data. Check the **Binary** box. Then **Min value** and **Max value** fields will automatically be set to 0 and 1, respectively.

Octal data. Check the **Octal** box. Then **Code length** field will be available to specify a length of the **Octal Code** input values, if None is selected in the **Differential Set** field. For example (also see §2.2), if an underlying differential set consists of 15 differentials, the corresponding **Octal Code length** should be 5, and the **Octal Code** data should be strings of 5 digits from the set of eight integers {0, 1, 2, 3, 4, 5, 6, 7} (e.g. 35164, 13704, 34520, 02250 etc). The **Code length** is mandatory parameter. Note that an **Octal data** table will possess only a single column of codes; so the **Number of columns** is irrelevant parameter, and the corresponding field will be disabled.

Hexadecimal data. Check the **Hexadecimal** box. Then **Code length** field will be available to specify a length of the **Hexadecimal Code** input values, if None is selected in the **Differential Set** field. For example (also see §2.2), if an underlying differential set consists of 24 differentials, the corresponding **Hexadecimal Code length** should be 6, and the **Hexadecimal Code** data should be strings of 6 letters from the set of the first sixteen consonants of English alphabet {B, C, D, F, G, H, J, K, L, M, N, P, R, S, Q, T}

(e.g. BNJHGR, DDTGLK, TTTTTT etc). The **Code length** is mandatory parameter. Note that a **Hexadecimal data** table will possess only a single column of codes; so the **Number of columns** is irrelevant parameter, and the corresponding field will be disabled.

Binary-Decimal data. Check the **Binary-Decimal** box. Then **Binary vector length** field will be available to specify a length of the **Binary vectors** corresponding to the **Binary-Decimal Code** input values, if None is selected in the **Differential Set** field. For example (also see §2.2), if an underlying differential set consists of 4 differentials, the corresponding **Binary vector length** should be 4, and the **Binary-Decimal Code** data should be $2^4 = 16$ integers from 0 to 15, representing the following sixteen binary vectors: $0 \leftrightarrow (0,0,0,0)$, $1 \leftrightarrow (0,0,0,1)$, $2 \leftrightarrow (0,0,1,0)$, $3 \leftrightarrow (0,0,1,1)$, ..., $15 \leftrightarrow (1,1,1,1)$. The **Binary vector length** is mandatory parameter, and its value must not exceed 63 ($2^{64} - 1$ is the maximum integer allowed by an operational system of PC). Note that a **Binary-Decimal data** table will possess only a single column of codes; so the **Number of columns** is irrelevant parameter, and the corresponding field will be disabled.

Number of rows (No. rows). It is mandatory to enter number of rows (individuals) in a new data table.

Number of columns (No. cols). It is mandatory to enter number of columns (differentials) in a new data table if none of the existing differential sets is selected (None is chosen in the **Differential Set** field).

There are two buttons **Back** and **Next** at the bottom of **Data Parameters** window (Fig. 2.13). Click **Back** to return to the main **Data entry** window. If values of all mandatory parameters have already been selected or entered, by clicking **Next**, a new window with two sheets **Original Data** (Fig. 2.14; see §2.3) and **Comments** will appear. The **Original Data** sheet (Fig. 2.14; also see §2.3) is composed of an empty grid (table) with dimension according to the specified data parameters. Comments about the data can be entered in the **Comments** sheet.

The **General Management Bar** (can be found above most VAT windows) contains **File**, **Change Application** and **Help** options. Under **File** there is the **Exit** option which allows user to leave the program and terminate current VAT session. **Change Application** provides a switch to other **VAT Applications** or return to the **VAT Main Window**. The **Help** function (under development) is for information about

VAT. In the **Original Data** sheet the **General Management Bar** contains an important additional option, namely **Edit**, as well as an **Excel** button.

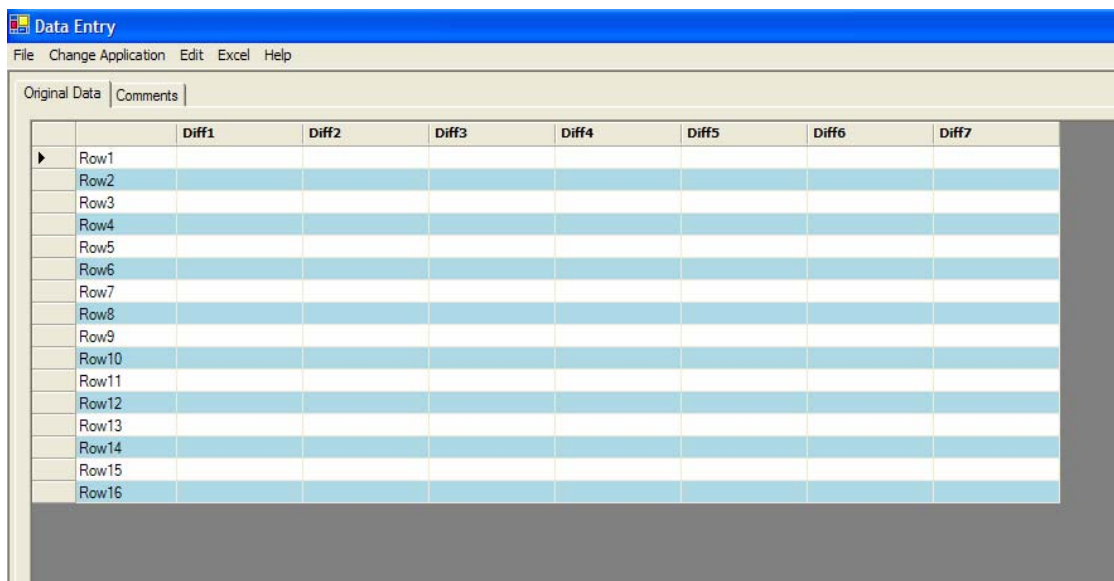


Figure 2.14: The empty grid (table) that was created by clicking **Next** in the **Enter New Data** window. In accordance with Fig. 2.13, there are 7 columns (named Diff1, Diff2 ...), and 16 rows.

2.5 Open Existing File

Open Existing File tool provides a possibility to import data from already existing file (the system **vat-file** or the text **txt-file**). This file can be uploaded and is available for modification. By clicking the **Open Existing File** square in the **Data Entry** window (Fig. 2.1), new window **Open existing file** (Fig. 2.15) will appear.

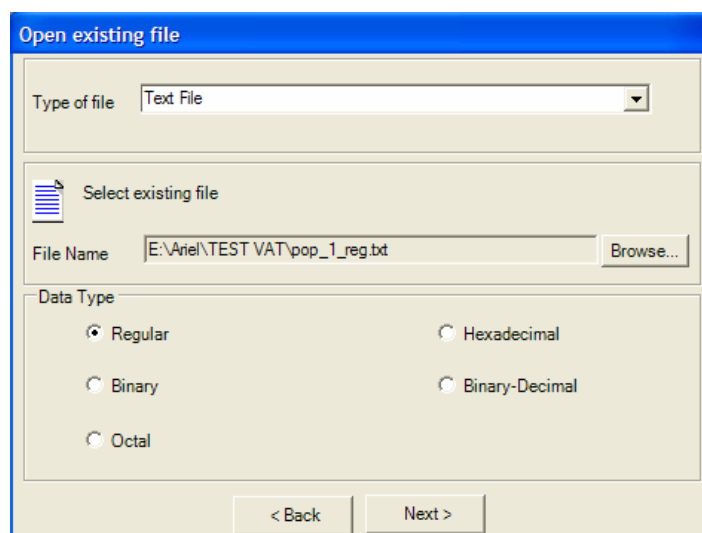


Figure 2.15: **Open existing file** window. The selected **Type of file** is Text File. The **Regular data** are declared to be imported from the selected existing file E:\Ariel\TEST VAT\pop_1_reg.txt.

Two mandatory parameters must always be determined in the **Open existing file** window: **Type of File** and **File name**. In the case of the text **txt-files**, the third parameter **Data Type** is also mandatory.

Type of File. One of the two possible file types, System VAT File or Text File, should be selected in the corresponding field. **No other** types of files can be imported.

Open existing System VAT File (vat-file). This type of file is the system-oriented file that is acceptable and can be used in all applications of the program. Once the System VAT type of file chosen, an input file must be determined using the standard browsing tools by clicking the **Browse** button. Only folders and available **vat-files** will appear. Once the file is found and entered in the **File name** field, click the **Next** button at the bottom of the **Open existing file** window. The **Original Data** sheet (Fig. 2.3) will be opened and the requested data appear in the grid. Clicking the **Back** button at the bottom will return to the **Data Entry** window.

Open existing Text File (txt-file). This type of file is acceptable and can be used only in the **Data Entry** applications. Once the Text type of file chosen, an input file must be selected using the standard browsing tools by clicking the **Browse** button. Only folders and available **txt-files** will appear. Once the file is found and entered in the **File name** field, the **Data Type** must be determined. Click proper circle to select among the five listed data types: **Regular**, **Binary**, **Octal**, **Hexadecimal** or **Binary-Decimal**. A wrongly specified type of data will result in an “invalid data” error message at the next stages. Click **Next** to continue. (Clicking the **Back** button at the bottom will return to the **Data Entry** window). Clicking **Next** will open a new window **Preview of file content (unformatted)** (Fig. 2.16). This window displays the data of the selected txt-file in an unformatted way. This representation allows recognizing if the data contain comments and names of columns and rows (a column delimiter can also be recognized). In order to upload the text file properly, the corresponding parameters should be specified in the field "Number of lines reserved for comments" and by clicking in the boxes "First column is reserved for names of rows" and/or "First row is reserved for names of columns" if there are a header row and/or a header column, respectively, in the given text file. Once choosing the appropriate parameters, continue by clicking **Next**. (Clicking **Back** will return to the **Open Existing File** window (Fig. 2.15)).

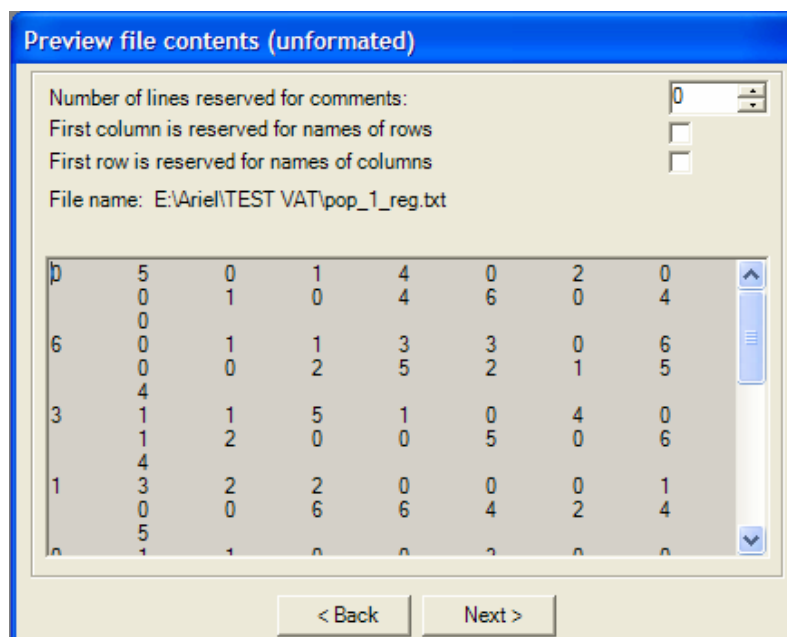
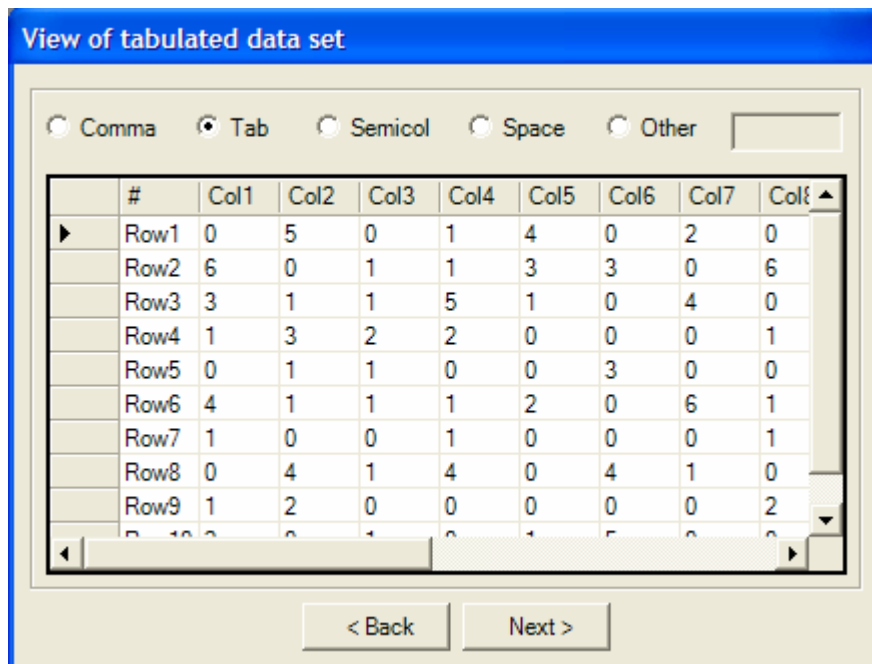


Figure 2.16: Preview of file content (unformatted) window displays **Regular data** delimited by tab, with no comments, and no names of rows and columns.

The next window is the **View of Tabulated Data Set** one (Fig. 2.17). This window displays the input data in a grid according to the parameters determined in the previous **Preview of file** window (Fig. 2.16) and an automatically recognized column delimiter used in the given txt-file. The VAT is able to recognize four commonly used standard delimiters: Comma, Tab, Semicolon, and Space. These delimiters together with an option "Other" are listed above the grid. The VAT-recognized delimiter (if it is identified) is automatically marked at the circle next to the delimiter name. If the delimiter was correctly recognized and the parameters were accurately determined in the previous **Preview of file** window, then the input data properly appear in the grid. By clicking the **Next** button at the bottom of the **View of Tabulated Data Set** window, the **Original Data** sheet (Fig. 2.3) will be opened, and the requested data will appear in the grid after automatic testing on validity according to the declared Data Type in the **Open Existing File** window (Fig. 2.15). In the case of invalid data, an error message will pop-up.

The view of data may be inappropriate if program could not find the correct delimiter and/or wrong parameters were determined in the **Preview of file** window (Fig. 2.16). One can return to the **Preview of file** window (click **Back**) to change the parameters. In order to fit the data, another column delimiter can be specified either among the optional standard ones or using the "Other" option (click in the circle next to

the corresponding option). By choosing "Other", any delimiter can be entered into the adjacent field. Once choosing the delimiter, click the "Go" button to get a new view of data. If a proper data appearance in the **View of Tabulated Data Set** window cannot be reached by means of the proposed changes of parameters and delimiters, then the only way of importing correct original data is to modify the data format in the requested Text File (**txt-file**).



a

		Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Col13	Col14	Col15	Col16
▶	Row1	0	5	0	1	4	0	2	0	0	1	0	4	6	0	4	0
	Row2	6	0	1	1	3	3	0	6	0	0	2	5	2	1	5	4
	Row3	3	1	1	5	1	0	4	0	1	2	0	0	5	0	6	4
	Row4	1	3	2	2	0	0	0	1	0	0	6	6	4	2	4	5
	Row5	0	1	1	0	0	3	0	0	0	1	0	4	0	2	6	1
	Row6	4	1	1	1	2	0	6	1	0	5	0	5	5	1	2	4
	Row7	1	0	0	1	0	0	0	1	1	2	4	4	5	1	4	4
	Row8	0	4	1	4	0	4	1	0	1	0	0	4	1	2	5	6
	Row9	1	2	0	0	0	0	0	2	0	1	1	5	3	0	1	2
	Row10	2	0	1	0	1	5	0	0	0	0	3	1	6	0	5	5

b

Figure 2.17: View of tabulated data set window: (a) View of **Regular data** from the existing Text File. The delimiter "Tab" was automatically recognized. (b) The data appearance in the **Original Data** sheet (Fig. 2.3).

§3. Resampling and Coding

This VAT application allows performing four operations, namely

- 1) To assign a cut-off and transform **Regular** into **Binary data (Conversion)**.
- 2) To transform **Octal**, **Hexadecimal** and **Binary-Decimal** data into **Binary data (Conversion)**.
- 3) To translate a **Binary data** set to all possible codes (**Octal**, **Hexadecimal** and **Binary-Decimal**) and one type of encoded data to another (**Coding**).
- 4) To draw random samples with replacement from a given data set (**Resampling**) for **Inferential Statistics**.

The **Conversion** of non-binary (**Regular**, **Octal**, **Hexadecimal** and **Binary-Decimal**) to **Binary data** (see §3.1) is the first mandatory and crucial task of the **Resampling and Coding** section. This conversion step is an essential precondition for further analysis, since the **Descriptive Statistics** as well as the **Inferential Statistics** applications both can only operate on binary data sets. **Conversion Regular data** is conducted as soon as a user-defined cut-off value is devised.

The **Coding** function allows viewing the original input data, as well as the converted binary data. Furthermore, it displays the **Octal**, **Hexadecimal** and **Binary-Decimal** representations (nomenclatures) of the **Binary data**.

The **Resampling** function generates a user-defined **number of fictive samples**, where rows are randomly drawn with replacement from a **Binary data** set. The **size of the computer-generated fictive samples** is also user-defined. Notice that the **Inferential Statistics** tool relies completely on these collections of randomly resampled sets.

3.1 The Files Selection Sheet

By clicking the **Resampling and Coding** square in the **VAT Main window**, a new window will appear with one open sheet called **Files selection**. This sheet is subdivided into three parts (Fig. 3.1), two segments (left and middle, named **source** and **work segment**, respectively) and right of these two segments a list of parameter values.

The **General Management Bar** (found above most VAT windows) contains the options **File**, **Change Application** and **Help**. Under **File** there is the **Exit** option which allows leaving the program and terminating current VAT session. With the **Change Application** option one can switch to other sections of the program or return to the **VAT Main Window**. The **Help** button provides information how to work with VAT.

The **source segment** on the left (with legend “All valid and complete system files”) provides a list of data files to select from for further analysis. The small path-line field (labeled **Folder**) above the **source segment** displays the full path of the **active folder** which holds the files listed in the **source segment**. Note that exclusively **vat**-files are listed, since they are the only ones to be processable in this section.

To select another folder click the **Browse** button next to the path-line field and use the browse option; the listed **vat**-files in the **source segment** will change accordingly.

Conversion by Transfer

One may **transfer** some *specific* files separately (by clicking “>”) or *all* listed files (by clicking “>>”) from the left **source segment** into the **work segment** (middle). Transfer only those files which have actually to be processed in this section.

The **work segment** accepts only **binary** data. If any non-binary data file was selected for transfer, it will automatically be converted first into binary before being transferred from the left (**source segment**) to the **work segment** (middle). For **Regular data** a user-defined **Cut-Off Value** is requested in a window that will pop up at the stage of a file **transfer** from the **source segment** into the **work segment**.

The “**Cut-Off Value**”-window (Fig. 3.2) is mainly comprised of the two boxes “**Replace with 0**” and “**Replace with 1**”, preceded by the name of the data file to be converted. The program automatically reads the minimum and maximum values from this data file and displays them under the headings “**From Min**” and “**To Max**”, respectively. If a **Cut-Off Value** has already been provided in an earlier session, this value will be displayed in the box labeled “**To Cut-Off**” (and a message at the bottom will alert you about its existence). One may then either change the old value or leave it unchanged. If previously no value was ever provided, then the “**To Cut-Off**” box will be empty, and a desired **Cut-Off Value** must be entered into the box. Next click the “**Apply**”-button, which tells the program to convert all values up to the **Cut-Off** into 0’s, and all values above the **Cut-Off Value** into 1’s. Finally, invoke the conversion by clicking **OK**. As a consequence the name of the processed file will appear in the “work-window” and its conversion parameters will be listed on the right side of the **Files selection** sheet (Fig. 3.1).

Before transferring any selected file from the **source** to the **work segment** the program will step by step convert each non-binary file.

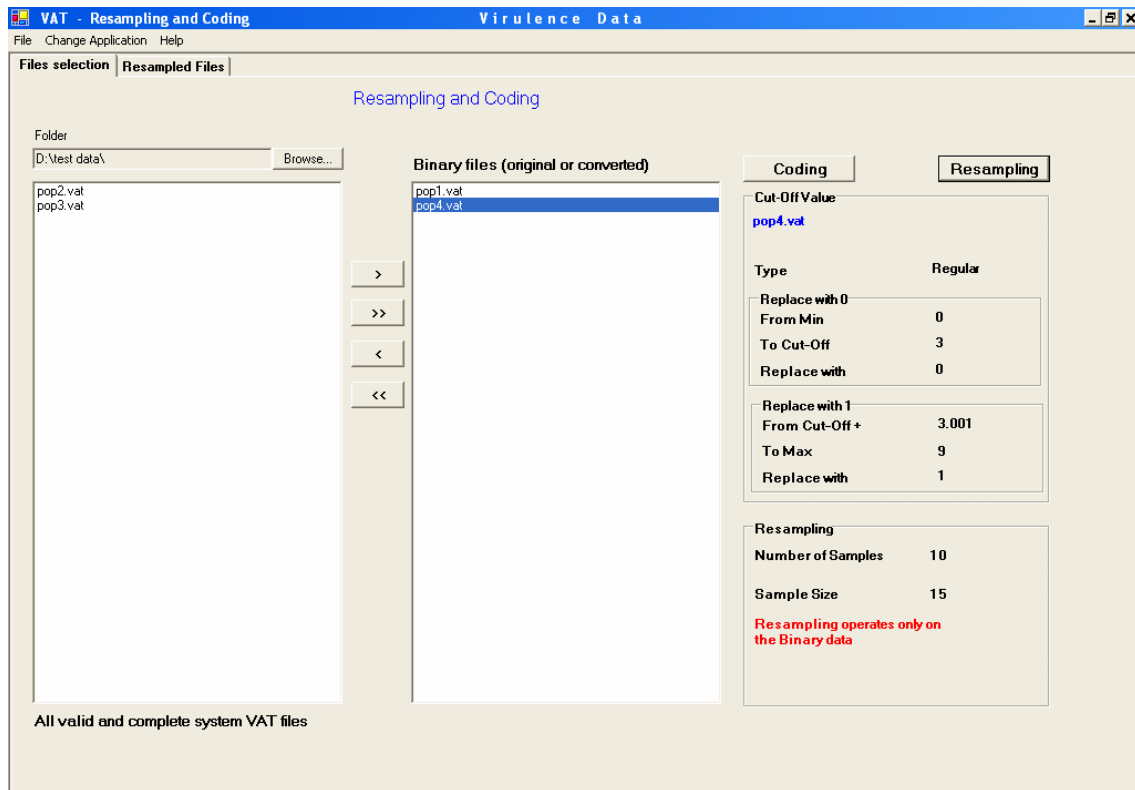


Figure 3.1: Files selection sheet in **Resampling and Coding**. The folder D:\test data\ contained four **vat**-files, two of which are listed in the **source segment** (left): pop2.vat and pop3.vat. Two files were transferred from the source to the **work segment** (middle): pop1.vat and pop4.vat. The parameters for the highlighted file “pop4.vat” with **Regular data** are listed on the right of the work segment, with Min = 0, Max = 9, and a **Cut-Off Value** of 3. Moreover, the two files in the work segment were resampled, namely 10 random samples of size 15 were drawn from each of the two files.

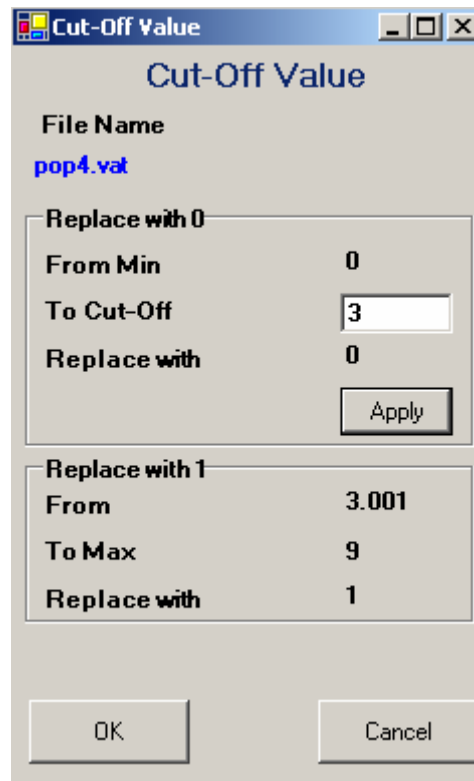


Figure 3.2: Cut-Off window. The Cut-Off value of the file pop4.vat was set to 3, then the “Apply”-button was clicked which automatically set the “From”-value in the “**Replace with 1**” to 3.001.

Once the specified files are successfully transferred to the **work segment**, two functions can be applied to them, namely **Coding** or **Resampling**. These functions appear as buttons on the top right side above the parameter list in the **Resampling and Coding** window (Fig. 3.1)

3.2 The Coding function

The “**Coding**”-function translates and displays the binary data in the three major nomenclatures, (1) Habgood’s **Binary-Decimal** code, (2) Gilmore’s **Octal** (triplet) code, and (3) Roelfs’ **Hexadecimal** code. To invoke this feature just click the “**Coding**”-button above the parameter list (Fig. 3.1) and all files in the “work-window” will be processed. Behind the **Files Selection** sheet four new sheets will appear labeled **Original input**, **Binary representation**, **Coded representation**, and **Comments**. Next the four new worksheets are explained in more detail.

The Original Input Sheet

This sheet shows the original input data matrix (Fig. 3.3). The small **Current file** field on top of the data table provides the name of the **vat**-file containing the displayed data.

The browse-option (click the downward arrow to the right of the **Current file** field) allows choosing between all files that had just been processed by the “**Coding**”-function.

Clicking the **Excel** button at the top will open an Excel-file and copy the contents of all four **Coding** sheets into four Excel-sheets.

	col1	col2	col3	col4	col5	col6	col7	col8	col9	col10	col11	col12	col13	col14	col15	col16
Row1	6	5	4	2	1	1	2	6	0	4	2	6	2	7	5	9
Row2	5	2	0	3	1	2	5	5	0	4	3	3	8	4	8	0
Row3	8	2	0	8	0	1	8	4	0	4	0	2	4	1	4	0
Row4	1	3	1	8	5	2	9	7	2	1	8	6	4	8	1	8
Row5	0	6	8	8	8	1	7	9	5	5	5	5	9	8	2	0
Row6	0	5	6	8	9	4	4	8	8	2	8	8	8	2	3	3
Row7	8	4	9	8	6	5	2	8	1	2	5	5	7	6	6	6
Row8	9	2	7	8	5	6	5	5	4	5	8	1	5	9	5	9
Row9	6	9	4	8	8	3	8	2	7	6	6	4	6	5	8	8
Row10	5	8	6	8	7	2	9	0	7	9	9	4	2	8	5	4
Row11	4	8	8	5	4	5	6	0	7	8	6	4	1	7	4	4
Row12	6	8	2	1	5	8	9	0	9	4	9	8	4	7	7	4

Regular No. cols: 16 No. rows: 12 Min value: 0 Max value: 9 Cut-Off Value: 3

Current file path D:\test data\pop4.vat

Figure 3.3: Original input sheet displaying the input data matrix of pop4.vat. Below the table all major data parameters are listed. Note: The data type here is **Regular**, and the **Cut-Off Value** is 3. This will cause all values below (above) 3 to become 0 (1, respectively) after conversion to binary form (for example, see col. 7, rows 1 and 2 in Fig. 3.4).

The Binary Representation Sheet

This sheet displays the binary data matrix after conversion (Fig. 3.4). For non-binary input data (i.e. **Regular**, **Binary-Decimal**, **Octal**, or **Hexadecimal**) a “**conversion by transfer**” took place; in the case of **Regular data** this conversion went according to a user-defined **Cut-Off Value**. For original input of type “**Binary**” the tables in the “**Original input**” and the “**Binary representation**” sheets coincide.

	col1	col2	col3	col4	col5	col6	col7	col8	col9	col10	col11	col12	col13	col14	col15	col16
Row1	1	1	1	0	0	0	0	1	0	1	0	1	0	1	1	1
Row2	1	0	0	0	0	0	1	1	0	1	0	0	1	1	1	0
Row3	1	0	0	1	0	0	1	1	0	1	0	0	1	0	1	0
Row4	0	0	0	1	1	0	1	1	0	0	1	1	1	1	0	1
Row5	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Row6	0	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
Row7	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1
Row8	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1
Row9	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1
Row10	1	1	1	1	1	0	1	0	1	1	1	1	1	0	1	1
Row11	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1
Row12	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1

Figure 3.4: Binary representation sheet displaying the data of pop4.vat after conversion to the binary form, applying a **Cut-Off Value** of 3 (compare with original **Regular** input in Fig. 3.3; for example, col. 7, rows 1 and 2).

Clicking the **Excel** button at the top will open an Excel-file and copy the contents of all four **Coding** sheets into four Excel-sheets.

The Coded Representation Sheet

This sheet shows the data in the three major nomenclatures, if applicable: the **Octal**, **Hexadecimal**, and **Decimal** codes. The **Octal** and **Hexadecimal** codes require suitable numbers of columns (multiples of three and four, respectively) and will only be displayed if these conditions are met. Note, that if the number of columns exceeds **63** then the Binary-Decimal code will not be calculated.

The screenshot shows the 'VAT - Resampling and Coding' application window. The 'Coded representation' tab is selected. The current file is 'pop4.vat'. The table displays 12 rows of data with columns for Name, Octal, Hexadecimal, and Binary-Decimal. The Octal column is empty for all rows. The Hexadecimal and Binary-Decimal columns contain codes for each row. The status bar at the bottom shows 'Regular' mode, 16 columns, 12 rows, min/max values of 0/9, and a cut-off value of 3. The current file path is 'D:\test data\pop4.vat'.

Name	Octal	Hexadecimal	Binary-Decimal
Row1		SCHK	57687
Row2		LF6S	33614
Row3		MFGN	37706
Row4		CPFR	6973
Row5		KPTQ	31740
Row6		KTPL	32696
Row7		TRFT	64831
Row8		PTST	49135
Row9		TNTT	64255
Row10		TNTK	64247
Row11		TSTK	65271
Row12		QSTT	52991

Regular No. cols : 16 No. rows : 12 Minvalue : 0 Maxvalue : 9 Cut-Off Value : 3
 Current file path D:\test data\pop4.vat

Figure 3.5: Coded representation sheet. Codes of rows in pop4.vat are represented. Only the **Hexadecimal** and **Binary-Decimal** codes are available. The **Octal** code cannot be calculated since 16 (number of columns) is no multiple of three, and the corresponding column is empty.

The screenshot shows the 'VAT - Resampling and Coding' application window. The 'Coded representation' tab is selected. The current file is 'p3.vat'. The table displays 21 rows of data with columns for Name, Octal, Hexadecimal, and Binary-Decimal. The Octal column contains codes for each row. The Hexadecimal and Binary-Decimal columns contain codes for each row. The status bar at the bottom shows 'Binary (0.1)' mode, 12 columns, 21 rows, and the current file path is 'D:\test VAT 16_02_08\p3.vat'.

Name	Octal	Hexadecimal	Binary-Decimal
Row1	6461	RFC	3377
Row2	5461	PFC	2865
Row3	5461	PFC	2865
Row4	4661	MPC	2481
Row5	4661	MPC	2481
Row6	4661	MPC	2481
Row7	4276	LPS	2238
Row8	4276	LPS	2238
Row9	4276	LPS	2238
Row10	4276	LPS	2238
Row11	6256	QNS	3246
Row12	6256	QNS	3246
Row13	6256	QNS	3246
Row14	6256	QNS	3246
Row15	6256	QNS	3246
Row16	0040	BDB	32
Row17	0040	BDB	32
Row18	0040	BDB	32
Row19	0040	BDB	32
Row20	0040	BDB	32
Row21	0040	BDB	32

Binary (0.1) No. cols : 12 No. rows : 21
 Current file path D:\test VAT 16_02_08\p3.vat

Figure 3.6: Coded representation sheet. Codes of rows in p3.vat are represented. Since the number of columns is 12 (it's a multiple of 3 as well as 4), hence **Octal** and **Hexadecimal** codes can both be obtained.

Clicking the **Excel** button at the top will open an Excel-file and copy the contents of all four **Coding** sheets into four Excel-sheets.

The Comments Sheet

This sheet displays the available comments about each analyzed population (file), these comments had to be included before in the **Data entry** section, see §2.4 and Fig.2.2.

Clicking the **Excel** button at the top will open an Excel-file and copy the contents of all four **Coding** sheets into four Excel-sheets.

Note: To view the data of a **vat**-file after **Coding** there are three sheets available (**Original input**, **Binary** and **Coded representation**). But only the **Current file** box in the **Original input** sheet allows choosing which file to view.

3.3 Resampling Function

The **Resampling** is necessary function for **Inferential Statistics** application (§5) which runs only with resampled binary data. The **Resampling** function takes a binary matrix from the **work segment** (Fig. 3.1) and generates a **number of fictive samples**, where rows are randomly drawn with replacement from the binary matrix. The **number** and **size of the computer-generated fictive samples** are user-defined. These samples based on the original data, provide a data pool which allows the assessment of statistical significance of population parameters and other statistical inferences in the **Inferential Statistics** application.

Once clicking the **Resampling** button at the top right, a **Resampling** window will pop up (Fig. 3.7). In this window the **Number of samples** that will be generated (default is 100) and the **Sample size** (number of rows in each sample) must be determined. Note that the scale of the resampled data correlates with the level of significance of the statistical analysis in the **Inferential Statistics** section (§5). On the other hand, large samples and a high number of samples lead to longer processing times. This is especially critical when running several files under the **Between-population** analysis (§5.3) where completing this analysis could take up to several hours.

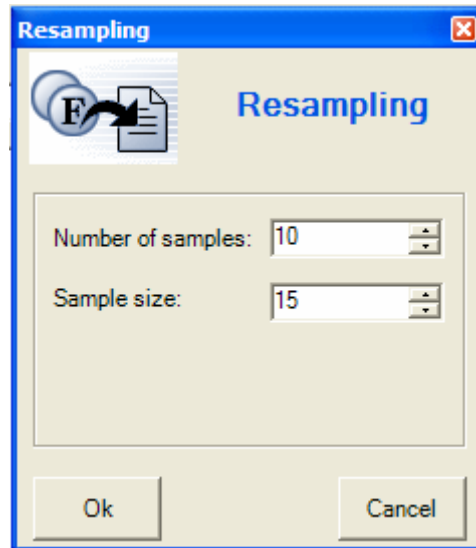


Figure 3.7: Resampling window. Number of samples is set to 10, and Sample size is set to 15.

When the **Resampling** parameters are determined, click **OK** in the **Resampling** window. **Resampling with these parameters will be done for all files in the work segment (middle)**. If one of the files was already resampled in a previous analysis, then a message will appear, and a user is required to choose if to run **Resampling** again under the new parameters. After **Resampling** is accomplished, a label of new sheet **Resampled Files** will appear. Clicking the **Resampled Files** will display a list of all processed files and resampling parameters **Number of samples** and **Sample size** (Fig. 3.8).

VAT - Resampling and Coding

Virulence Data

File

Change Application

Help

Files selection

Resampled Files

List of Resampled Files

	File	Number of samples	Sample size
▶	pop2.vat	10	15
	pop3.vat	10	15
	pop1.vat	10	15
	pop4.vat	10	15

Figure 3.8: Resampled Files sheet. Four files pop1.vat, pop2.vat, pop3.vat and pop4.vat were resampled (the **Number of samples** and the **Sample size** are 10 and 15, respectively).

Once **Resampling** is run, the **Resampling** parameters will be displayed at the right bottom corner of the **Files Selection** sheet in the **Resampling and Coding** window (Fig. 3.1).

Clicking the **Excel** button at the top will open an Excel-file and copy the **List of Resampled files** into Excel-sheet.

§4. Descriptive Statistics

This VAT application allows you to characterize individual populations (within) or groups of population (between) by purely descriptive statistical method. You can analyze all those data that were previously prepared with the applications **Data Entry** (§2) and **Conversion** (§3.1) and are available as vat-type files. (**No other files** will be accepted for the following descriptive analysis procedures.)

It is possible to perform a **Within**-analysis of single populations or a **Between**-analysis of pairs of populations. Inferential Statistics can be performed in the corresponding application (see §5), however only after Resampling (see §3).

Note: Formulae of all parameters and indices calculated by the VAT, brief explanations and the list of relevant literature can be found in **Appendix**. The Appendix formulae are designated A1, A2 etc, and the corresponding references to them appear in the manual.

4.1 The Files Selection Sheet

By clicking the **Descriptive statistics** square in the **VAT Main window**, a new window will appear with one open sheet called **Files selection**. This sheet is subdivided into three parts (Fig. 4.1), two segments (left and middle, named **source** and **work segment**, respectively) and right of these two segments a list of parameter values.

The **General Management Bar** (found above most VAT windows) contains the options **File**, **Change Application** and **Help**. Under **File** there is the **Exit** option which allows leaving the program and terminating current VAT session. With the **Change Application** option one can switch to other sections of the program or return to the **VAT Main Window**. The **Help** button provides information how to work with VAT.

The **source segment** on the left (with legend “Binary files original or converted”) provides a list of data files to select from for further analysis. The small path-line field (labeled **Folder**) above the source segment displays the full path of the **active folder** which holds the files listed in the **source segment**. Note that exclusively **vat**-files are listed, since they are the only ones to be processable in this section.

To select another folder click the **Browse** button next to the path-line field and use the browse option; the listed **vat**-files in the **source segment** will change accordingly.

One may **transfer** some *specific* files separately (by clicking “>”) or *all* listed files (by clicking “>>”) from the left **source segment** into the **work segment** (middle). Transfer only those files which have actually to be processed in this section.

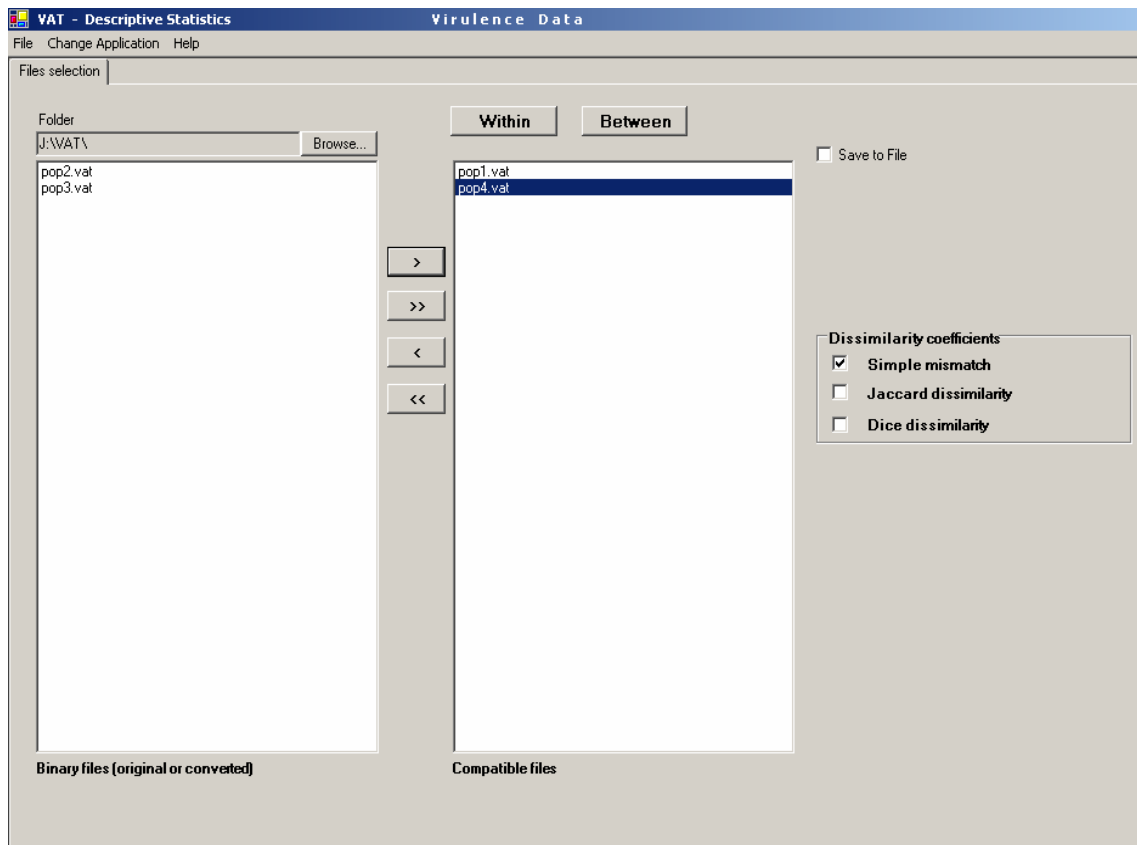


Figure 4.1: Files selection sheet in the **Descriptive Statistics** section. Two files, pop1.vat and pop4.vat, were transferred to the **work segment** (middle) from the J:\ VAT\ folder. The rest of the folder’s files, pop2.vat and pop3.vat, are displayed in the **source segment** (left). Only **vat**-files with already existing **Binary** representation of original data are displayed. On the right there is a list of **Dissimilarity coefficients** where the **Simple mismatch** box is checked.

Save to File. Before running the **Between** or **Within** options one can choose to save the results as a **text file**. In order to do this check the **Save to File** box on the right and select a folder and file name. After running the **Between** or **Within** the results of calculations will be saved in the corresponding **txt-file**.

Once the desired files are in the **work segment** (middle), two types of population analyses can be activated: **Within** or **Between**.

Within includes a series of population analyses that are applied separately to each file (population) selected in the **work segment**. By clicking the **Within** button all these analyses will be performed. The **Within** population analysis is further explained in §4.2.

Between includes a series of analyses for pairwise comparison of populations (files). In the case of more than two files, all possible pairs will be analyzed. **Note that the Between procedure can only be run if the number and names of columns are identical for all data tables in the files selected in the work segment.** Otherwise an error message will appear. If the number of rows is not identical in all selected data, files the Kosman's *KBm*, *KBj* and *KBd* (equation A31 in **Appendix**) cannot not be calculated. The program will still run. The **Between** population analysis is further explained in §4.3.

Before running the **Between** or **Within** population analysis one should choose among three **Dissimilarity coefficients**. For that purpose check one or more of the associated boxes on the right, labeled **Simple mismatch** (A1), **Jaccard dissimilarity** (A2) and **Dice dissimilarity** (A3). (Note, that A1, A2 etc are designations of equations that appear in **Appendix**).

4.2 Within-population analysis

Within includes a series of population analyses that are applied separately to each file (population) selected in the **work segment**. By clicking the **Within** button all these analyses will be performed.

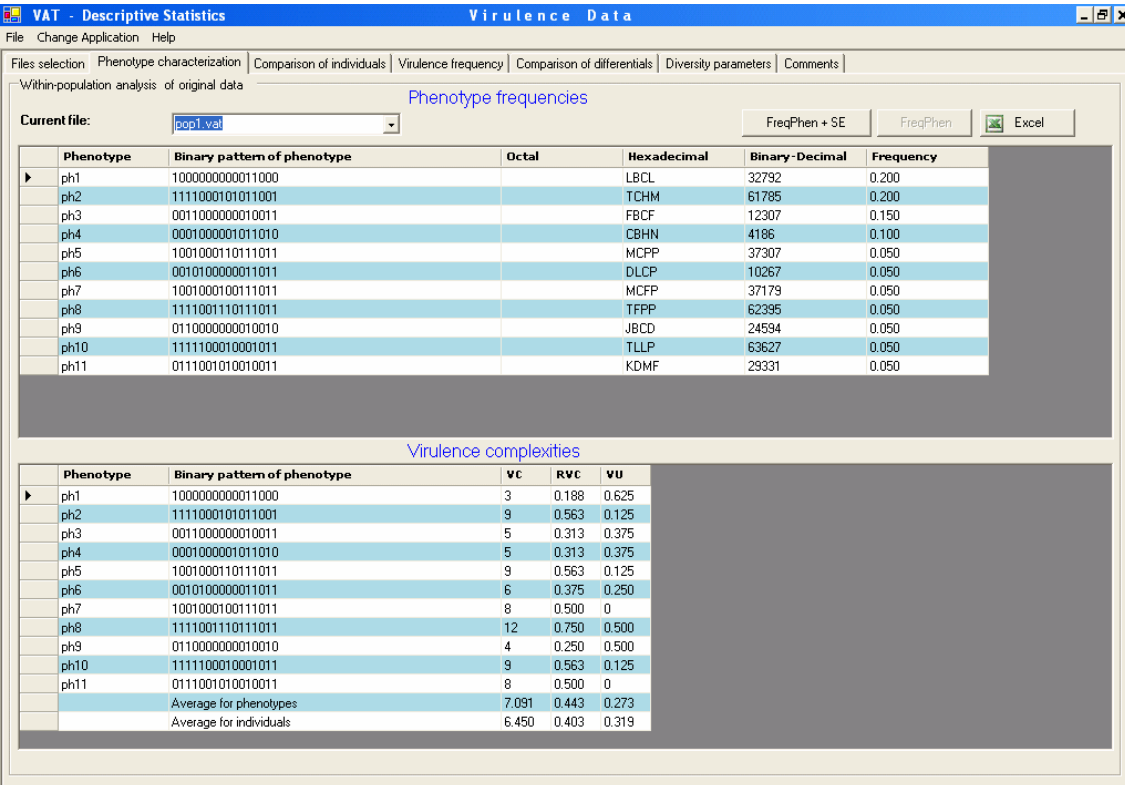
Once **Within** is activated, it could take a few minutes. At the bottom of the window the file name currently analyzed and the elapsed running time are displayed as well as a blue bar conveying graphically the approximate time still required for completion. As soon as the analysis is completed, six new sheets will appear behind the **Files selection** sheet, only the name tags of the six new worksheets will be visible right above the upper edge of the front sheet. Clicking such name tag brings the corresponding sheet to the foreground, in Fig. 4.2, for example, **Phenotype characterization** is the visible front worksheet, while **Files selection** and the other five sheets **Comparison of Individuals**, **Virulence frequency**, **Comparison of differentials**, **Diversity parameters** and **Comments** are in the back. Next we explain the new worksheets in more detail.

The Phenotype Characterization Sheet

This sheet shows in two tables the analysis of **Phenotype frequency** and **Virulence complexities** (see Fig. 4.2). In the **current file's** field above the table one can select from all analyzed files. On the right to this field are three buttons: **FreqPhen+SE**, **FreqPhen**, and **Excel**.

While **FreqPhen** provides the proportion of each phenotype in the underlying sample (A18), **FreqPhen + SE** provides also the standard error **SE** (A19), calculated under the Binomial model.

Clicking the **Excel** button will open an Excel-file and copy the contents of both tables from the VAT sheet into Excel-sheets.



The screenshot shows the 'VAT - Descriptive Statistics' window with the 'Virulence Data' tab selected. The 'Current file' is set to 'pop1.vat'. The 'Phenotype frequencies' table (upper) lists 11 phenotypes (ph1 to ph11) with their binary patterns, Octal, Hexadecimal, Binary-Decimal, and Frequency values. The 'Virulence complexities' table (bottom) lists the same 11 phenotypes with their VC, RVC, and VU values. The 'FreqPhen' button is selected in the top right corner.

Phenotype	Binary pattern of phenotype	Octal	Hexadecimal	Binary-Decimal	Frequency
ph1	100000000011000		LBCL	32792	0.200
ph2	1111000101011001		TCHM	61785	0.200
ph3	0011000000010011		FBCF	12307	0.150
ph4	0001000001011010		CBHN	4186	0.100
ph5	1001000110111011		MCPP	37307	0.050
ph6	0010100000011011		DLCP	10267	0.050
ph7	1001000100111011		MCFP	37179	0.050
ph8	1111001110111011		TFFP	62395	0.050
ph9	0110000000010010		JBCD	24594	0.050
ph10	1111100010001011		TLLP	63627	0.050
ph11	011001010010011		KDMF	29331	0.050

Phenotype	Binary pattern of phenotype	VC	RVC	VU
ph1	100000000011000	3	0.188	0.625
ph2	1111000101011001	9	0.563	0.125
ph3	0011000000010011	5	0.313	0.375
ph4	0001000001011010	5	0.313	0.375
ph5	1001000110111011	9	0.563	0.125
ph6	0010100000011011	6	0.375	0.250
ph7	1001000100111011	8	0.500	0
ph8	1111001110111011	12	0.750	0.500
ph9	0110000000010010	4	0.250	0.500
ph10	1111100010001011	9	0.563	0.125
ph11	011001010010011	8	0.500	0
Average for phenotypes		7.091	0.443	0.273
Average for individuals		6.450	0.403	0.319

Figure 4.2: Phenotype characterization window. The Phenotype frequency table (upper table) contains list of 5 different phenotypes revealed in the chosen file (pop1.vat) with their parameters. These phenotypes are also analyzed in the Virulence complexities table (bottom table). At the top right corner the **FreqPhen** option is chosen.

The Phenotype Frequencies Table

This table lists all phenotypes found in the underlying sample. The first column assigns names to each phenotype (ph1, ph2, ...). Since each isolate in the sample is expressed by a binary vector, isolates with identical binary vectors are defined to exhibit

the same phenotype. The second column displays for each phenotype the associated binary vector. **Recall** that the dimension of the vector (its “length”) equals the number of underlying differentials. The next three columns contain the corresponding **Octal**, **Hexadecimal**, and **Binary-Decimal** race codes (see §2.2). The last column of the Phenotype frequencies table shows the **Phenotype Frequency** (A18) with or without standard error **SE** (A19), depending on the choice between the **FreqPhen + SE** or **FreqPhen** button.

The Three Implemented Race Codes

Octal (or triplet) code (see §2.2) requires a binary vector of length divisible by three; otherwise the **VAT** leaves the Octal column empty.

Hexadecimal code (see §2.2) requires a binary vector of length divisible by four; otherwise the **VAT** leaves the Hexadecimal column empty. (See also the **VATView-tool**).

Binary-Decimal code (see §2.2) is available only for binary vectors up to a length of 63; otherwise the **VAT** leaves the Binary-Decimal column empty.

The Virulence Complexities Table

The first two columns from the left are the same as in the **Phenotype Frequencies** table: On the left is a list of all the different phenotypes, the second column displays the associated binary vectors. The next three columns, labeled **VC**, **RVC**, and **VU**, contain values of the **Virulence Complexity** (A6), the **Relative Virulence Complexity** (A7), the **Virulence Uniformity** (A8), respectively. The last two rows of the **Virulence Complexities** table contain the average values of **VC**, **RVC** and **VU**, calculated over all phenotypes and over all individuals, respectively.

The Comparison of Individuals Sheet

This sheet displays all pairwise dissimilarities between individuals (isolates) of the population (Fig. 4.3) in a form of a lower triangular matrix. By checking the **Square** box on the top right the dissimilarities will be displayed in a symmetric square matrix.

The **Current file** field allows selecting among all analyzed files.

The **Dissimilarity measures** field allows choosing from all measures that have already be selected for analysis before in the **Files selection** sheet (see Fig. 4.1, right side). **VAT** allows the choice among the following three commonly used measures: **Simple mismatch** (A1), **Jaccard dissimilarity** (A2), and **Dice dissimilarity** (A3).

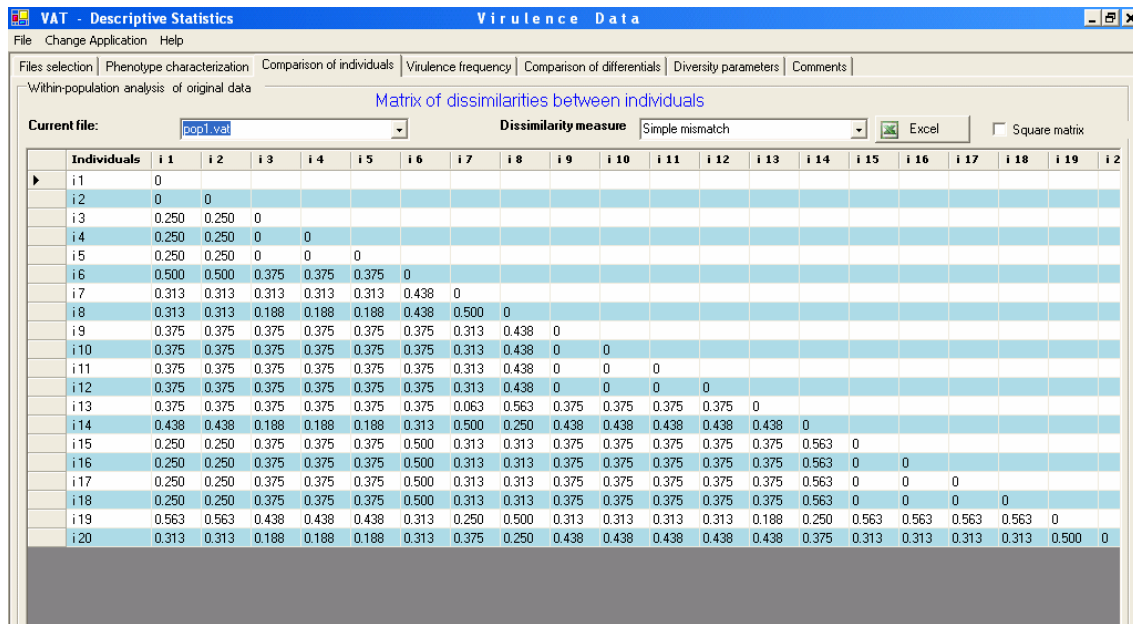


Figure 4.3: Comparison of individuals sheet. Here all the 20 individuals of a population (pop1.vat) are compared pair wise between themselves and the **Simple mismatch** measure is selected. The matrix is a lower triangle matrix (square box not checked).

Clicking the **Excel** button at the top right will open an Excel-file and copy the contents into an Excel-sheet.

The Virulence Frequency Sheet

This sheet displays proportion of the sampled isolates with virulent reaction on each differential from the given differential set (Fig. 4.4). (In other words, the frequency of 1s in the corresponding column of the original data table is displayed). The leftmost column lists all members (host plants) of the underlying differential set, labeled by default as Col1, Col2, etc. The first two rows serve the table's header and provide the designations and file names of the analyzed populations. Values of virulence frequencies alone or together with the Standard Error (**SE**) can be displayed by clicking the appropriate button (**FreqVir** or **FreqVir +SE**, respectively).

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

VAT - Descriptive Statistics

Virulence Data

File Change Application Help

Files selection | Phenotype characterization | Comparison of individuals | **Virulence frequency** | Comparison of differentials | Diversity parameters | Comments

Within-population analysis of original data

Virulence frequency

FreqVir+SE FreqVir Excel

	P1	P2	P3	P4
Differential	pop1.vat	pop4.vat	pop2.vat	pop3.vat
Col1	0.600 ± 0.110	0.750 ± 0.125	0.667 ± 0.136	0.367 ± 0.122
Col2	0.400 ± 0.110	0.667 ± 0.136	0.667 ± 0.136	0.367 ± 0.122
Col3	0.600 ± 0.110	0.667 ± 0.136	0.583 ± 0.142	0.400 ± 0.126
Col4	0.700 ± 0.102	0.750 ± 0.125	1 ± 0.000	0.733 ± 0.114
Col5	0.100 ± 0.067	0.750 ± 0.125	0 ± 0.000	0 ± 0.000
Col6	0 ± 0.000	0.417 ± 0.142	0.083 ± 0.080	0.367 ± 0.064
Col7	0.100 ± 0.067	0.833 ± 0.108	0.417 ± 0.142	0.333 ± 0.122
Col8	0.350 ± 0.107	0.667 ± 0.136	0.500 ± 0.144	0.400 ± 0.126
Col9	0.200 ± 0.089	0.583 ± 0.142	0.500 ± 0.144	0.400 ± 0.126
Col10	0.300 ± 0.102	0.750 ± 0.125	0 ± 0.000	0 ± 0.000
Col11	0.150 ± 0.080	0.750 ± 0.125	0.583 ± 0.142	0.467 ± 0.129
Col12	0.950 ± 0.049	0.750 ± 0.125	1 ± 0.000	1 ± 0.000
Col13	0.750 ± 0.097	0.750 ± 0.125	0.750 ± 0.125	0.300 ± 0.103
Col14	0 ± 0.000	0.833 ± 0.108	0 ± 0.000	0 ± 0.000
Col15	0.600 ± 0.110	0.750 ± 0.125	0.417 ± 0.142	0.333 ± 0.122
Col16	0.650 ± 0.107	0.667 ± 0.136	1 ± 0.000	1 ± 0.000

Figure 4.4: Virulence frequency sheet. Isolates from four populations (pop1.vat, pop2.vat, pop3.vat, pop4.vat) were tested on the set of 16 differentials Col1, Col2,..., Col6 (columns in the original data table). Here the FreqVir+SE option is chosen, so the virulence frequencies together with the corresponding values of the standard error are displayed.

The Comparison of Differentials Sheet

This sheet displays **Associations** (A5) or **Correlations** (A4) between virulences and avirulences for all pairs of the differentials (columns in the original data table) from the given differential set (see Fig. 4.5) in form of an upper or lower triangular matrix, respectively. By checking the **Square** box on the top the corresponding triangular matrix will be displayed in a symmetric square matrix. The option **Both** allows combining the **Associations** and **Correlations** results (upper and lower triangular matrix, respectively) in one square matrix. The **Current file** field allows selecting among all analyzed files.

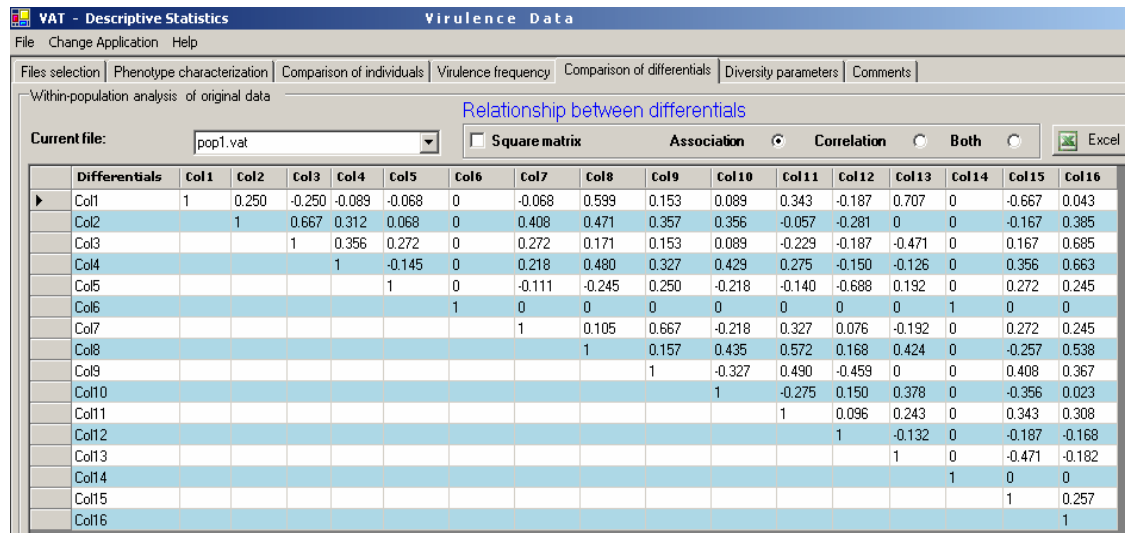


Figure 4.5: Comparison of differentials sheet. Virulence/avirulence reactions of isolates to 16 differentials are analyzed with respect to a given population (pop1.vat file). The **Association** measure is selected and displayed in an upper triangular matrix. (The **Square matrix** box is unchecked here.)

The Diversity Parameters Sheet

This sheet provides different measures of diversity within populations namely **Nei Diversity (Hs, A27)**, **Simpson (Si, A21)**, **Normalized Shannon (Sh, A25)**, **Kosman Index (K, A29)**, **Stoddart (St, A22)**, **Shannon (SH, A23)**, **Evenness (E, A24)**, **Gleason (G, A20)**, **Average dissimilarity within (ADW, A16)**, **Kosman diversity within (KW, A16)**.

The Indices **ADW** and **KW** can be calculated with regard to the three commonly used dissimilarity measures: **Simple Mismatch (m , A1)**, **Jaccard (j , A2)**, and **Dice (d , A3)**.

Values of all relevant diversity parameters are displayed in a table where each column corresponds to one of the analyzed populations.

Clicking the **Excel** button at the top right will open an Excel-file and copy the contents into an Excel-worksheet.

The Comments Sheet

This sheet displays the available comments about each analyzed population, these comments had to be included before in the **Data entry** section, see §2.4 and Fig.2.4.

4.3 Between-Populations Analysis

Clicking the **Between** button on the top right in the **Files selection** sheet (see §4.1 and Fig. 4.1) of the **Descriptive Analysis** section activates several procedures for

comparisons between the selected data sets (populations) as listed by their file names in the **work segment** (central part of sheet).

Note: In order to be compatible those data need to be based on identical differential sets, i. e. **the number and names of columns in the original data tables must be equal in all selected files**. For the indices **KBm**, **KBj**, and **KBd** the number of rows must also be equal in all these files.

Once **Between** is activated, it could run for several minutes. The bottom line reports about the current file in process and about the elapsed running time. A blue bar conveys graphically the approximate time still required for completion. As soon as the analysis is completed, several new sheets will appear behind the **Files selection** sheet, only the name tags of these six new worksheets will be visible right above the upper edge of the front sheet. Clicking such name tag brings the corresponding sheet to the foreground, in Fig. 4.6, for example, **Pairwise common phenotype** is the visible front worksheet, while **Files selection** and the other five sheets **Comparison of phenotypes**, **Overall common phenotypes**, **Virulence frequency**, **Distance parameters between samples**, **Comments** are in the back. Next we explain the new worksheets in more detail.

The Pairwise Common Phenotype Sheet

This sheet displays for any selected pair of populations the phenotypes common to both. The **current pair of files** field at the top allows you to select among the analyzed file pairs. Once a pair is chosen, a table is displayed with the **binary** phenotype vectors, the corresponding **Octal**, **Hexadecimal**, and **Binary-Decimal** (see §2.2) codes of each of the common phenotypes. Note that only the common phenotypes are shown.

The two columns **Count** (**CountP2** and **CountP3** in Fig. 4.6) of the table report the number of individuals with a common phenotype encountered in the respective population. For example, the third phenotype #3 (Fig. 4.6), which is found in both populations **P2** and **P3**, appears 2 times in the first population **P2**; therefore, in the corresponding cell of phenotype #3 and **CountP2**, the number 2 is displayed. The column **Mincounts** gives the minimum value 1 of **CountP2** (=2) and **CountP3** (=1) for phenotype #3.

The two columns **Freq** (**FreqP2** and **FreqP3** in Fig. 4.6) contain the relative frequencies of the common phenotypes (A18) with respect to the sample size of the corresponding population. For example, the third phenotype #3 (Fig. 4.6) occurs 2 times

among a total of 12 isolates in sample **P2** and only once among a total of 15 isolates in sample **P3**; then **FreqP2** and **FreqP3** equal to $2/12=0.167$ and $1/15=0.067$, respectively.

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

#	Binary pattern of phenotype	Octal	Hexadecimal	Binary-Decimal	Count P2	Count P3	Min Counts	Freq P2	Freq P3
1	1111001110111011		TFPP	62395	3	3	3	0.250	0.200
2	0001000110111001		CCPM	4537	2	2	2	0.167	0.133
3	1111000000011001		TBCM	61465	2	1	1	0.167	0.067
4	0111000000011011		KBCP	28699	1	1	1	0.083	0.067
5	0001000000010001		CBCC	4113	1	1	1	0.083	0.067
6	1001010110111011		MHPP	38331	1	1	1	0.083	0.067
7	1101001000010001		RDCC	53777	1	1	1	0.083	0.067

Figure 4.6: Pairwise common phenotype sheet. The 7 common phenotypes are revealed and analyzed in pairwise comparison of two populations (pop2.vat and pop3.vat). **Note: for file names P1, P2, P3 ... see comments sheet.**

The Comparison of Phenotypes Sheet

This sheet displays dissimilarity between phenotypes of two populations (files) under comparison, and it is comprised of three tables (Fig. 4.7). The **Current pair of files** field at the top allows selecting among the analyzed file pairs.

The **Dissimilarities between phenotypes matrix** displays the pairwise dissimilarity values between phenotypes of the two populations. The field **Dissimilarity measure** at the top allows choosing from all those measures (**Simple Mismatch**, **Jaccard**, **Dice**) which were previously selected in the **Files selection** sheet (see Fig. 4.1, right side) for analysis. The **rows** are linked to the phenotypes of the **first** population (file name left of “+” in the **Current pair of files** field), and the **columns** to the **second** one.

The **Phenotype characterization** tables at the lower part of the sheet list all phenotypes for the current pair of populations. The first column assigns numbers to each phenotype (ph1, ph2, ...). The second column displays for each phenotype the associated binary vector. **Recall** that the dimension of the vector (its “length”) equals the number of underlying differentials. The next three columns contain the corresponding **Octal**, **Hexadecimal**, and **Binary-Decimal** race codes (see §2.2). Finally, the **Frequency** column gives the frequency of each phenotype.

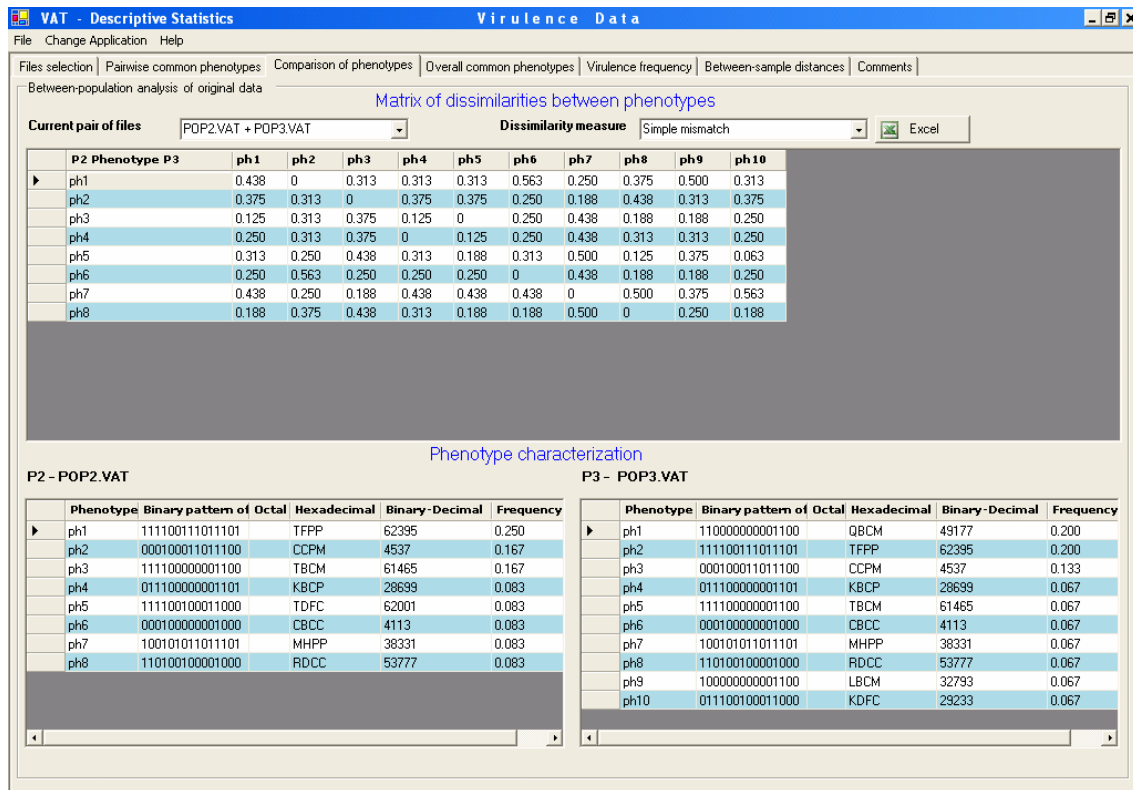


Figure 4.7: Comparison of phenotypes sheet. Two files are compared (pop2.vat and pop3.vat). The top matrix contains the dissimilarity values between the 8 phenotypes of pop2.vat and the 10 phenotypes of pop3.vat. At the bottom right and left are the **Phenotype characterization** of pop2.vat and pop3.vat, respectively.

The Overall Common Phenotypes Sheet

This sheet (Fig. 4.8) displays the phenotypes **common to all** analyzed files simultaneously (**not** just pairwise common as in Fig. 4.6).

The table lists the binary vector of each common phenotype, then in the next three columns gives the corresponding **Octal**, **Hexadecimal**, and **Binary-Decimal** codes (see §2.2) for each. The following columns **FreqP1**, **FreqP2**, **FreqP3**, ... show the frequencies of each phenotype in each population. For example, if population P1 has sample size 12 and a common phenotype is found 3 times in P1 then its **FreqP1** value is $3/12 = 0.25$. The column **Minfreq** gives the minimum value of **FreqP1**, **FreqP2**, **FreqP3**, ...

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

#	Binary pattern of phenotype	Octal	Hexadecimal	Binary-Decimal	Freq P1	Freq P2	Freq P3	Min Freq
1	1111001110111011		TFPP	62395	0.050	0.250	0.200	0.050

Figure 4.8: Overall common phenotype sheet. One common phenotypes was found in all 3 populations (pop1.vat, pop2.vat, pop3.vat).

The Virulence Frequency Sheet

This sheet displays proportion of the sampled isolates with virulent reaction on each differential from the given differential set (Fig. 4.9). (In other words, the frequency of 1s in the corresponding column of the original data table is displayed). The leftmost column lists all members (host plants) of the underlying differential set, labeled by default as Col1, Col2, etc. The first two rows serve the table's header and provide the designations and file names of the analyzed populations. Values of virulence frequencies alone or together with the Standard Error (SE) can be displayed by clicking the appropriate button (**FreqVir** or **FreqVir +SE**, respectively).

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

	P1	P2	P3
Differential	pop1.vat	pop2.vat	pop3.vat
Col1	0.600	0.667	0.667
Col2	0.400	0.667	0.667
Col3	0.600	0.583	0.400
Col4	0.700	1	0.733
Col5	0.100	0	0
Col6	0	0.083	0.067
Col7	0.100	0.417	0.333
Col8	0.350	0.500	0.400
Col9	0.200	0.500	0.400
Col10	0.300	0	0
Col11	0.150	0.583	0.467
Col12	0.950	1	1
Col13	0.750	0.750	0.800
Col14	0	0	0
Col15	0.600	0.417	0.333
Col16	0.650	1	1

Figure 4.9: Virulence frequency sheet. Virulence frequencies on the set of 16 differentials (columns) are analyzed with respect to 3 different populations (pop1.vat, pop2.vat, pop3.vat). At the top right the **FreqVir** option is chosen.

The Between-Sample Distance Sheet

Matrices of pairwise distances between the analyzed populations are provided. Seven different distance measures are available, namely **Nei distance N** (A38), **Nei's Gst** (A42), **Kosman's Gst** (A43), **Rogers distance** (A32), **Mean Characters difference MCD** (A44), **DADm** (A30), **Kosman distance KBm** (A31).

For each measure a matrix in a separate sheet can be accessed by clicking the corresponding sheet name tag visible at the upper edge of the front sheet, Fig. 4.10.

By checking the **Square** box on the top the corresponding triangular matrix will be displayed in a symmetric square matrix.

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

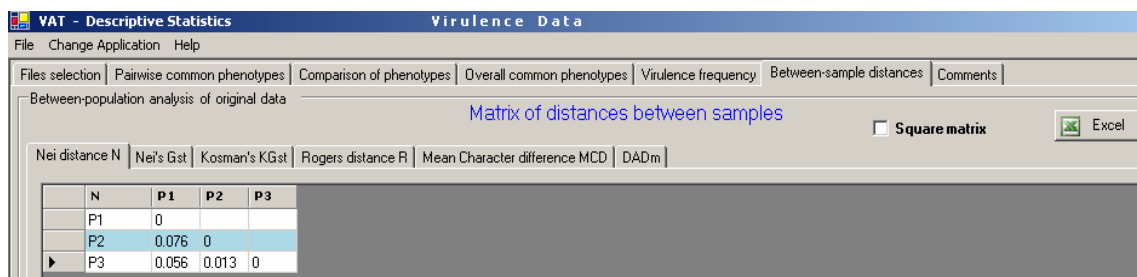


Figure 4.10: Between-sample Distances sheet. The Nei distance N sheet is selected, and three populations are compared (pop1.vat, pop2.vat, pop3.vat). The **Square matrix** box is not checked.

The Comments Sheet

This sheet (Fig. 4.11) displays the available comments about each analyzed population, these comments had to be included before in the **Data entry** section, see §2.4 and Fig.2.4. Since some tables (e.g. Fig. 4.10) use general labels like **P1**, **P2**, ... for the analyzed populations (files), the **Comment sheet** may be useful to learn which file is associated to each label.

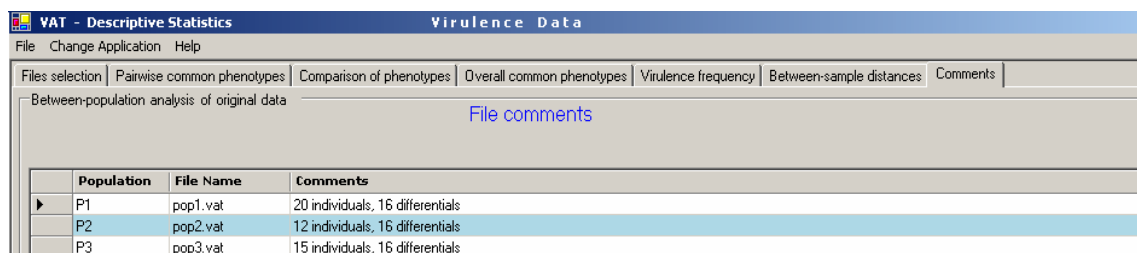


Figure 4.11: Comments sheet. Three different populations (including file names) are shown (pop1.vat, pop2.vat, pop3.vat).

§5. Inferential Statistics

This VAT application allows statistically analyzing individual populations (**Within**) or groups of population (**Between**), and also estimating statistical significance of calculated parameters based on resampled data.

Only resampled binary files (original or converted) obtained in the **Resampling and Coding** application (see §3) are acceptable and can be read in the **Inferential Statistics** application.

One can analyze **Within** a single file (population) or **Between** different files (populations). This application is similar to the **Descriptive Statistics** application. The main difference between the **Inferential Statistics** and the **Descriptive Analysis** is that the former application provides **Average** values of parameters (A45) and the corresponding values of the **Standard Error SE** (A46) on the basis of the resampled data.

5.1 The Files Selection Sheet

By clicking the **Inferential Statistics** square in the **VAT Main window**, a new window will appear with one open sheet called **Files selection**. This sheet is subdivided into three parts (Fig. 5.1), two segments (left and middle, named **source** and **work segment**, respectively) and right of these two segments a list of parameter values.

The **General Management Bar** (found above most VAT windows) contains the options **File**, **Change Application** and **Help**. Under **File** there is the **Exit** option which allows leaving the program and terminating current VAT session. With the **Change Application** option one can switch to other sections of the program or return to the **VAT Main Window**. The **Help** button provides information how to work with VAT.

The **source segment** on the left (with legend “Binary files after resampling”) provides a list of data files to select from for further analysis. The small path-line field (labeled **Folder**) above the source segment displays the full path of the **active folder** which holds the files listed in the **source segment**. Note that exclusively **vat**-files after **Resampling** (§3.3) are listed, since they are the only ones to be processable in this section. To select another folder click the **Browse** button next to the path-line field and use the browse option; the listed **vat**-files in the **source segment** will change accordingly.

Now one may **transfer** some *specific* files separately (by clicking “>”) or *all* listed files (by clicking “>>”) from the left **source segment** into the **work segment** (middle, with legend “Compatible files”). Transfer only those files into the **work segment** which are actually wanted to be processed in the current session.

The **Inferential Statistics** can **only be run** if the **number** and **names of columns** are **identical** in the **original data** tables **for all files** listed in the **work segment**. It is also required that the collections of re-sampled sets are compatible, i.e. the **number** of computer-generated samples as well as their **sample sizes** must be the same (see Fig. 3.7). Otherwise an error message will appear.

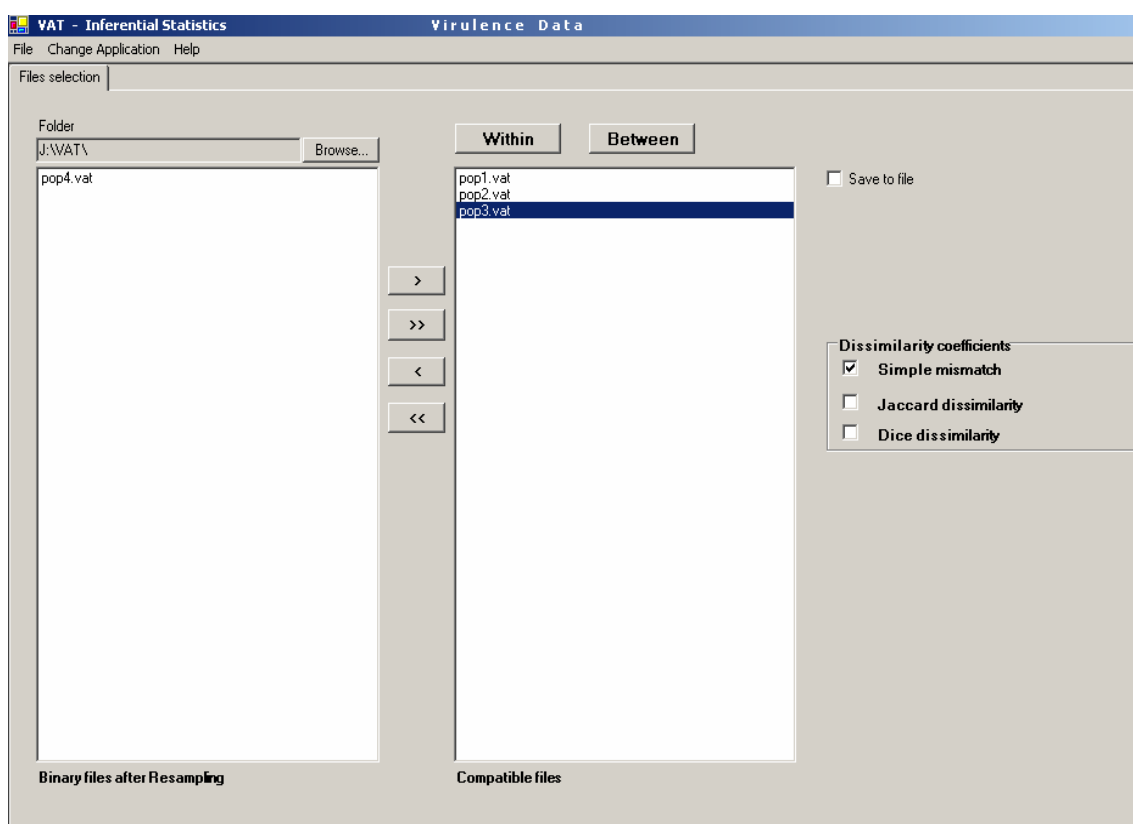


Figure 5.1: File selection sheet in the **Inferential Statistics** section. Three files were transferred to the **work segment** (middle) from the J:\VAT\ folder (pop1.vat, pop2.vat, pop3.vat). The rest of the folder’s suitable files are displayed on the left in the **source segment** (pop4.vat). Only **Binary** and **Resampled vat**-files are displayed. On the right there is a list of **Dissimilarity coefficients** where the **Simple mismatch** box is checked.

Save to File. Before running the **Between** or **Within** options one can choose to save the results as a **text file**. In order to do this check the **Save to File** box on the right and select a folder and file name. After running the **Between** or **Within** the results of calculations will be saved in the corresponding **txt-file**.

Once the desired files are in the **work segment** (middle), two types of population analyses can be activated: **Within** or **Between**.

Within includes a series of population analyses that are applied separately to each file (population) selected in the **work segment**. By clicking the **Within** button all these analyses will be performed. The **Within** population analysis is further explained in §5.2.

Between includes a series of analyses for pairwise comparison of populations (files). If more than two files are selected, then all possible pairs will be analyzed. **Note that the Between procedure can only be run if the number and names of columns are identical for all data tables in the files selected in the work segment.** Otherwise an error message will appear. The **Between** procedure is further explained in §5.3.

Before running the **Between** or **Within** population analysis the **Dissimilarity coefficients** must be chosen. For that purpose check one or more of the associated boxes on the right, labeled **Simple mismatch** (A1), **Jaccard** (A2) and **Dice** (A3).

5.2 Within-Population Analysis

Within includes a series of population analyses applied separately to each population (file) listed in the **work segment**. The **Within** population analysis is based on the corresponding collections of previously resampled sets. By clicking the **Within** button this analysis will be performed.

Once **Within** is activated, it could take a few minutes. A special information window will open where the file name currently analyzed and the elapsed running time are displayed as well as a blue bar conveying graphically the approximate time still required for completion. As soon as the analysis is completed, five new sheets will appear behind the **Files selection** sheet, only the name tags of the five new worksheets will be visible right above the upper edge of the front sheet. Clicking such name tag brings the corresponding sheet to the foreground, in Fig. 5.2, for example, **Phenotype characterization** is the visible front worksheet, while **Files selection** and the other four sheets **Virulence frequency**, **Comparison of differentials**, **Diversity parameters**, and **Comments** are in the back. Next we explain the new worksheets in more detail.

The Phenotype Characterization Sheet

This sheet shows in two tables the analysis of **Phenotype frequency** and **Virulence complexities** (see Fig. 5.2). In the **current file**'s field above the table you can select

from all analyzed files. On the right to this field are three buttons: **Average+SE**, **Average**, and **Excel**.

The **Average** provides the arithmetic means (A45), calculated over all the computer-generated random samples originating from the previous resampling runs (see §3.3).

The **Average+SE** provides the mean (A45) and the standard error **SE**(A46) of the estimated parameters.

Clicking the **Excel** button will open an Excel-file and copy the contents of both tables from the VAT sheet into Excel-sheets.

We describe now the two tables of the **Phenotype Characterization** Sheet in more detail.

Phenotype frequencies

Phenotype	Binary pattern of phenotype	Octal	Hexadecimal	Binary-Decimal	Frequency
ph1	100000000011000		LBCL	32792	0.253
ph2	1111000101011001		TCHM	61785	0.200
ph3	001100000010011		FBCF	12307	0.133
ph4	0001000001011010		CBHN	4186	0.087
ph5	1001000110111011		MCPP	37307	0.053
ph6	0010100000011011		DLCP	10267	0.080
ph7	1001000100111011		MCFP	37179	0.040
ph8	1111001110111011		TFPP	62395	0.047
ph9	011000000010010		JBCD	24594	0.047
ph10	1111100010001011		TLLP	63627	0.040
ph11	0111001010010011		KDMF	29331	0.020

Virulence complexities

Original Data

Phenotyp	Binary pattern of phenotype	VC	RVC	VU
ph1	100000000011000	3	0.188	0.625
ph2	1111000101011001	9	0.563	0.125
ph3	001100000010011	5	0.313	0.375
ph4	0001000001011010	5	0.313	0.375
ph5	1001000110111011	9	0.563	0.125
ph6	0010100000011011	6	0.375	0.250
ph7	1001000100111011	8	0.500	0
ph8	1111001110111011	12	0.750	0.500
ph9	011000000010010	4	0.250	0.500
ph10	1111100010001011	9	0.563	0.125
ph11	0111001010010011	8	0.500	0
Average for phenotypes		7.0909	0.443	0.273
Average for individuals		6.45	0.403	0.319

Resampled Data

Resampling estimates	VC	RVC	VU
Average for phenotypes	6.761	0.423	0.299
Average for individuals	6.207	0.388	0.344

Figure 5.2: Phenotype characterization sheet of **Inferential Statistics**. The Phenotype frequency table (upper table) contains list of 11 different phenotypes that were revealed and analyzed for the chosen population (file pop1.vat). The possible codes and the frequency of each phenotype are displayed. Note that the Octal representation is impossible for the set of 16 differentials. The results for the same phenotypes also appear in the Virulence complexities table for original data (bottom left table). The results for the resampled data appear in the bottom right table.

The Phenotype Frequency Table

This table lists all phenotypes found in the underlying sample. The first column assigns names to each phenotype (ph1, ph2, ...). Since each isolate in the sample is expressed by a binary vector, isolates with identical binary vectors are defined to exhibit the same phenotype.

The second column displays for each phenotype the associated binary vector. **Recall** that the dimension of the vector (its “length”) equals the number of underlying differentials.

The next three columns contain the corresponding **Octal**, **Hexadecimal**, and **Binary-Decimal** race codes (see §2.2).

Note: The **Octal** (or triplet) code requires a binary vector of length divisible by **three**; otherwise the VAT leaves the **Octal** column empty. The **Hexadecimal** code requires a binary vector of length divisible by **four**; otherwise the VAT leaves the **Hexadecimal** column empty. The **Binary-Decimal** code can be calculated only for binary vectors up to a length of 63; otherwise the VAT leaves the **Binary-Decimal** column empty.

The last column of the Phenotype frequency table shows the **Average** (A45) or **Average+SE** (A45-A46), depending on which button was activated (one may toggle between both buttons).

The Virulence Complexities Table

The left part of the **Virulence Complexities** table is labeled **Original Data** and contains exactly the same values as the corresponding table in **Descriptive Statistics** (see Fig. 4.2). The first two columns from the left are the same as in the **Phenotype Frequencies** table: On the left is a list of all the different phenotypes, the second column displays the associated binary vectors. The next three columns, labeled **VC**, **RVC**, and **VU**, contain values of the **Virulence Complexity** (A6), the **Relative Virulence Complexity** (A7), the **Virulence Uniformity** (A8), respectively. The last two rows of the **Virulence Complexities** table contain the average values of **VC**, **RVC** and **VU**, calculated over all phenotypes and over all individuals, respectively.

The second table in the lower right (see Fig.5.2) is labeled **Resampled Data** and contains the **Average for phenotypes** and **Average for individuals** for **VC**, **RVC**, and **VU**. However in contrast to the left **Original Data** table the **Resampled Data** table contains the average calculated over all computer-generated resampled data sets.

Depending on which button was activated (one may toggle between **Average** or **Average+SE**), only the **Average** (A45) or additionally the corresponding standard error **SE** (A46), respectively, will appear.

The Virulence Frequency Sheet

This sheet displays the resampling based estimates (A45-A46) of proportion of the sampled isolates with virulent reaction on each differential from the given differential set (Fig. 5.3). The leftmost column lists all members (host plants) of the underlying differential set, labeled by default as Col1, Col2, etc. The first two rows serve the table's header and provide the designations and file names of the analyzed populations. Average values (A45) of virulence frequencies alone or together with the Standard Error **SE** (A46) can be displayed by clicking the appropriate button (**Average** or **Average+SE**, respectively).

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

	P1	P2	P3
Differential	pop1.vat	pop2.vat	pop3.vat
Col1	0.633	0.640	0.627
Col2	0.353	0.640	0.613
Col3	0.567	0.533	0.420
Col4	0.620	1	0.793
Col5	0.120	0	0
Col6	0	0.073	0.060
Col7	0.067	0.420	0.327
Col8	0.340	0.500	0.427
Col9	0.160	0.500	0.427
Col10	0.287	0	0
Col11	0.140	0.593	0.487
Col12	0.960	1	1
Col13	0.800	0.720	0.787
Col14	0	0	0
Col15	0.547	0.367	0.327
Col16	0.613	1	1

Figure 5.3: Virulence frequency sheet. Virulence frequencies of the sampled isolates on the set of 16 differentials (columns) are calculated for two different populations P1 (pop1.vat), P2 (pop2.vat), P3 (pop3.vat). The Average option is chosen at the top right corner.

The Comparison of Differentials Sheet

This sheet is subdivided in two matrices (Fig. 5.4). The top matrix displays Average values (A45) of **Associations** (A5) or **Correlations** (A4) between virulences and avirulences for all pairs of the differentials (columns in the original data table) from the given differential set (see Fig. 5.4) in form of an upper or lower triangular matrix,

respectively. The bottom matrix displays the Standard Errors **SE** (A46) of the corresponding parameters. By checking the **Square** box on the top the corresponding triangular matrices will be displayed in a symmetric square ones. The option **Both** allows you to combine the upper and lower triangular matrices in one square matrix. The **Current file** field allows you to select among all analyzed files.

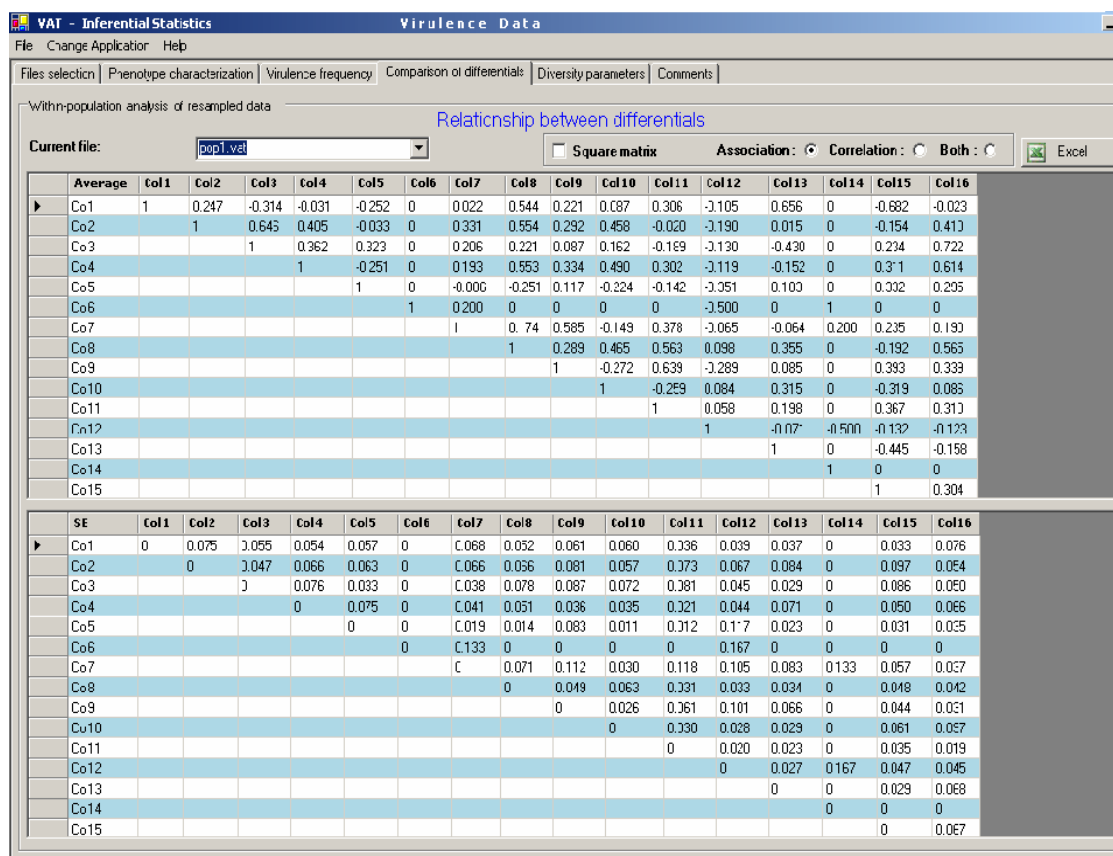


Figure 5.4: Comparison of differentials sheet. Virulence/avirulence reactions of isolates to 16 differentials are analyzed with respect to a given population (pop1.vat file). The **Association** measure is selected and displayed in an upper triangular matrix. (The **Square matrix** box is unchecked here.) The top and bottom matrices display the Average, based on the resampled data.

The Diversity Parameters Sheet

This sheet provides the resampling based estimates (A45-A46) of different measures of diversity within populations namely **Nei Diversity** (**Hs**, A27), **Simpson** (**Si**, A21), **Normalized Shannon** (**Sh**, A25), **Kosman Index** (**K**, A29), **Stoddart** (**St**, A22), **Shannon** (**SH**, A23), **Evenness** (**E**, A24), **Gleason** (**G**, A20), **Average dissimilarity within** (**ADW**, A16), **Kosman diversity within** (**KW**, A16).

The Indices **ADW** and **KW** can be calculated with regard to the three commonly used dissimilarity measures: **Simple Mismatch** (m , A1), **Jaccard** (j , A2), and **Dice** (d , A3).

Values of all relevant diversity parameters are displayed in a table where each column corresponds to one of the analyzed populations. Average values (A45) of the parameters alone or together with the Standard Error **SE** (A46) can be displayed by clicking the appropriate button (**Average** or **Average+SE**, respectively).

Clicking the **Excel** button at the top right will open an Excel-file and copy the contents into an Excel-worksheet.

The Comments Sheet

This sheet displays the available comments about each analyzed population, these comments had to be included before in the **Data entry** section, see §2.4 and Fig.2.14.

5.3 Between-populations analysis

Clicking the **Between** button on the top right in the **Files selection** sheet (see §5.1 and Fig. 5.1) of the **Inferential Statistics** section activates several procedures for comparisons between the selected data sets (populations) as listed by their file names in the **work segment** (central part of sheet).

Once **Between** is activated, it could run for several minutes or even hours. A special information window will open where the reports about the current pairs of files in process and about the elapsed running time are displayed as well as a blue bar conveying graphically the approximate time still required for completion. As soon as the analysis is completed, three new sheets will appear behind the **Files selection** sheet, only the name tags of these three new worksheets will be visible right above the upper edge of the front sheet. Clicking such name tag brings the corresponding sheet to the foreground, in Fig. 5.5, for example, **Virulence frequency** is the visible front worksheet, while **Files selection** and the other two sheets **Between population distances** and **Comments** are in the back.

Next we explain the three new worksheets in more detail.

The Virulence Frequency Sheet

This sheet displays the resampling based estimates (A45-A46) of proportion of the sampled isolates with virulent reaction on each differential from the given differential

set (Fig. 5.5). The leftmost column lists all members (host plants) of the underlying differential set, labeled by default as Col1, Col2, etc. The first two rows serve the table's header and provide the designations and file names of the analyzed populations. Average values (A45) of virulence frequencies alone or together with the Standard Error **SE** (A46) can be displayed by clicking the appropriate button (**Average** or **Average+SE**, respectively).

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

	P1	P2	P3
Differential	pop1.vat	pop2.vat	pop3.vat
Col1	0.633	0.640	0.627
Col2	0.353	0.640	0.613
Col3	0.567	0.533	0.420
Col4	0.620	1.000	0.793
Col5	0.120	0.000	0.000
Col6	0.000	0.073	0.060
Col7	0.067	0.420	0.327
Col8	0.340	0.500	0.427
Col9	0.160	0.500	0.427
Col10	0.287	0.000	0.000
Col11	0.140	0.593	0.487
Col12	0.960	1.000	1.000
Col13	0.800	0.720	0.787
Col14	0.000	0.000	0.000
Col15	0.547	0.367	0.327
Col16	0.613	1.000	1.000

Figure 5.5: Virulence frequency sheet. Virulence frequencies of the sampled isolates on the set of 16 differentials (columns) are calculated for two different populations P1 (pop1.vat), P2 (pop2.vat) and P3 (pop3.vat). The Average option is chosen at the top right corner.

The Between-Population Distances Sheet

Matrices of the resampling based estimates (A45-A46) of pairwise distances between the analyzed populations are provided. Seven different distance measures are available, namely **Nei distance N** (A38), **Nei's Gst** (A42), **Kosman's Gst** (A43), **Rogers distance** (A32), **Mean Characters difference MCD** (A44), **DADm** (A30), **Kosman distance KBm** (A31).

For each measure a separate sub-sheet can be accessed by clicking the corresponding **name tag**, which is visible at the upper edge of the front sheet, see Fig. 5.6. These index-specific sub-sheets are subdivided into two matrices. The top and bottom matrices display the Average values (A45) of the distance measure and the

corresponding Standard Errors **SE** (A46), respectively. All values are based on the corresponding **resampling data sets**.

By checking the **Square** box on the top the corresponding triangular distance and **SE** matrices will be displayed in the form of symmetric square ones.

Clicking the **Excel** button at the top right opens an Excel-file and copies the contents into an Excel-sheet.

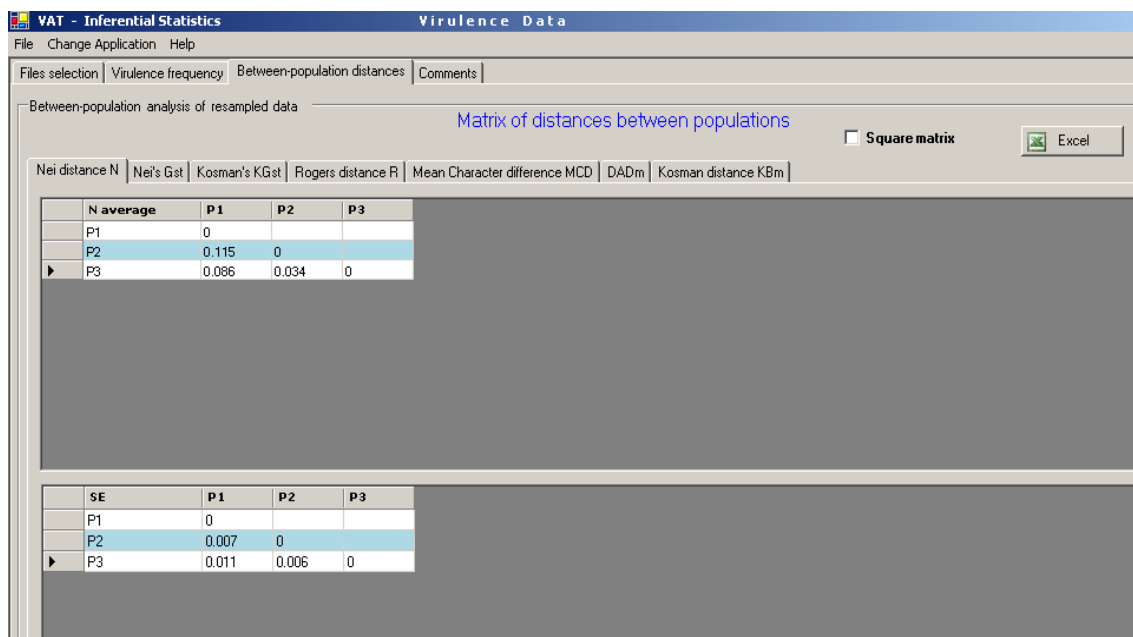
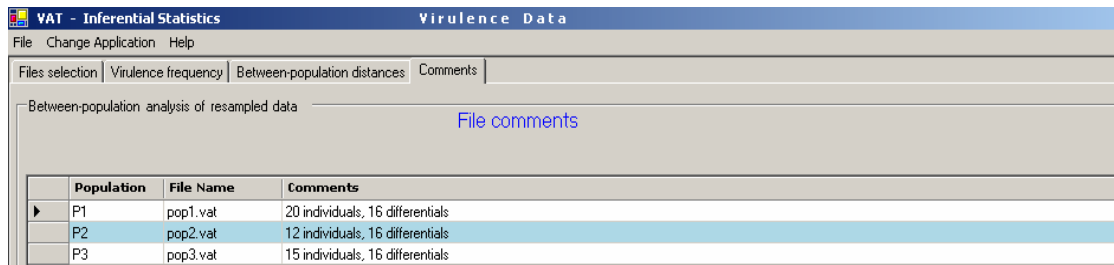


Figure 5.6: Between-population Distances sheet. The Nei distance N sheet is displayed, and three populations are compared (pop1.vat, pop2.vat, pop3.vat). The **Square matrix** box is not checked. The top and bottom matrices display the Average values, respectively, of distances between populations (look in **Comments** sheet for the file name of each population P1, P2 etc.).

The Comments Sheet

This sheet (Fig. 5.7) displays the available comments about each analyzed population. These comments had to be included before in the **Data entry** section, see §2.4 and Fig.2.4. Since some tables use general labels like **P1**, **P2** etc. for the analyzed populations (files), the **Comment sheet** may be useful to learn which file is associated to each label.



Population	File Name	Comments
P1	pop1.vat	20 individuals, 16 differentials
P2	pop2.vat	12 individuals, 16 differentials
P3	pop3.vat	15 individuals, 16 differentials

Figure 5.7: Comments sheet. Three populations and their file names are displayed with comments available.

Part II: Resistance Analysis applications

Data entry and computational tools for **Resistance Analysis** are nearly identical to those for **Virulence Analysis** described in detail in **Part I**. There are mainly differences in terminology and designations, where "Resistance" (R) is used instead "Virulence" (V). For example, **Resistance Complexity (CR)**, **Relative Resistance Complexity (RCR)**, **Resistance uniformity (UR)**, and **Resistance Frequency** in the **Resistance Analysis** correspond to **Virulence Complexity (VC)**, **Relative Virulence Complexity (RVC)**, **Virulence Uniformity (VU)**, and **Virulence Frequency** in the **Virulence Analysis**, respectively.

System files for **Resistance Data** have extension **rat** (filename.rat) similar to the extension **vat** (filename.vat) for the **Virulence Data**.

The most significant differences between **Resistance Analysis** and **Virulence Analysis** appear in the **Resampling and Coding** application.

§6. Resampling and Coding for Resistance Data

When an individual is tested for resistance to any pathogen isolate then the fact of resistance is usually designated by 0, i. e. the isolate is avirulent on the given individual. Susceptibility of an individual to any given isolate of pathogen is designated by 1 in the case of **Binary Data** (virulent reaction of the isolate) or by any number according to an assessment scale in the case of **Regular Data**. Therefore, in the binary form of resistance data 0 rather than 1 is a "valuable characteristics" what is unusual. To avoid this in the case of **Resistance Analysis**, the **Binary representation** sheet (Fig. 7.2; see also §3.2 and compare with Fig. 3.4 for **Virulence Analysis**) displays original data (Fig. 7.1) after conversion to the binary form and additional 0-1 transformation, so that in the

Binary representation sheet 1 already corresponds to resistance while 0 means susceptibility of the individual to the given isolate. The **Coded representation** sheet (Fig. 7.3) displays **Octal**, **Hexadecimal** and **Binary-Decimal** codes of the binary resistant patterns of individuals that appear in the **Binary representation** sheet (Fig. 7.2).

	col1	col2	col3	col4	col5	col6	col7	col8	col9	col10	col11	col12
Row1	1	1	0	1	0	0	1	1	0	0	0	1
Row2	1	0	1	1	0	0	1	1	0	0	0	1
Row3	1	0	1	1	0	0	1	1	0	0	0	1
Row4	1	0	0	1	1	0	1	1	0	0	0	1
Row5	1	0	0	1	1	0	1	1	0	0	0	1
Row6	1	0	0	1	1	0	1	1	0	0	0	1
Row7	1	0	0	0	1	0	1	1	1	1	1	0
Row8	1	0	0	0	1	0	1	1	1	1	1	0
Row9	1	0	0	0	1	0	1	1	1	1	1	0
Row10	1	0	0	0	1	0	1	1	1	1	1	0
Row11	1	1	0	0	1	0	1	0	1	1	1	0
Row12	1	1	0	0	1	0	1	0	1	1	1	0
Row13	1	1	0	0	1	0	1	0	1	1	1	0
Row14	1	1	0	0	1	0	1	0	1	1	1	0
Row15	1	1	0	0	1	0	1	0	1	1	1	0

Figure 7.1: Original input sheet displaying the input **Resistance Data** matrix of p1.rat.

Current file: p1.rat

	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12
Row1	0	0	1	0	1	1	0	0	1	1	1	0
Row2	0	1	0	0	1	1	0	0	1	1	1	0
Row3	0	1	0	0	1	1	0	0	1	1	1	0
Row4	0	1	1	0	0	1	0	0	1	1	1	0
Row5	0	1	1	0	0	1	0	0	1	1	1	0
Row6	0	1	1	0	0	1	0	0	1	1	1	0
Row7	0	1	1	1	0	1	0	0	0	0	0	1
Row8	0	1	1	1	0	1	0	0	0	0	0	1
Row9	0	1	1	1	0	1	0	0	0	0	0	1
Row10	0	1	1	1	0	1	0	0	0	0	0	1
Row11	0	0	1	1	0	1	0	1	0	0	0	1
Row12	0	0	1	1	0	1	0	1	0	0	0	1
Row13	0	0	1	1	0	1	0	1	0	0	0	1
Row14	0	0	1	1	0	1	0	1	0	0	0	1
Row15	0	0	1	1	0	1	0	1	0	0	0	1

Binary {0,1} No. cols : 12 No. rows : 15
Current file path: D:\test VAT 16_02_08\p1.rat

Figure 7.2: Binary representation sheet displaying the original **Resistance Data** of p1.rat (Fig. 7.1) after conversion to the binary form including the 0-1 transformation.

Current file: p1.rat

	Octal	Hexadecimal	Binary-Decimal
Row1	1316	DQS	718
Row2	2316	GQS	1230
Row3	2316	GQS	1230
Row4	3116	JGS	1614
Row5	3116	JGS	1614
Row6	3116	JGS	1614
Row7	3501	KGC	1857
Row8	3501	KGC	1857
Row9	3501	KGC	1857
Row10	3501	KGC	1857
Row11	1521	FHC	849
Row12	1521	FHC	849
Row13	1521	FHC	849
Row14	1521	FHC	849
Row15	1521	FHC	849

Binary {0,1} No. cols : 12 No. rows : 15
Current file path: D:\test VAT 16_02_08\p1.rat

Figure 7.3: Coded representation sheet. Codes of rows in the **Binary representation** sheet of the **Resistance Data** of p1.rat (Fig. 7.2) are represented.

For example, if original **Resistance Data** are of **Binary** type, then the following correspondence between binary original and transformed vectors and codes exists.

Table 1. Resistance Analysis

Original data	Binary representation	Coded representation		
		Octal	Hexadecimal	Binary-Decimal
111000101010	000111010101	0725	CSH	469

For comparison, the same binary vector in the case of original **Virulence Data** has the following representations.

Table 2. Virulence Analysis

Original data	Binary representation	Coded representation		
		Octal	Hexadecimal	Binary-Decimal
111000101010	111000101010	7052	QDN	3626

Similar to the **Virulence Analysis**, the **Binary representation** of original data are in fact used in the **Descriptive Statistics** and **Inferential Statistics** applications in the case of analyzing **Resistance Data**.

§7. Appendix

1. Dissimilarity between individuals.

Let us consider binary patterns ($\{0,1\}$ -vectors) of two individuals x and y tested on k differentiating factors D_s , $s = 1, 2, \dots, k$. We denote a = number of factors with shared positive responses (1s) for the both individuals, b = number of factors where individual x has a positive response, but y does not, c = number of factors where individual y has a positive response, but x does not.

m - Simple mismatch dissimilarity.

Simple mismatch coefficient of dissimilarity between two individuals x and y is determined as follows:

$$m(x, y) = \frac{b + c}{k}. \quad (\text{A1})$$

The simple mismatch coefficient varies between 0 and 1.

j - Jaccard dissimilarity.

Jaccard dissimilarity between two individuals x and y is determined as follows:

$$j(x, y) = \frac{b + c}{a + b + c}. \quad (\text{A2})$$

The Jaccard coefficient of dissimilarity varies between 0 and 1.

Note: $j(x, y)$ is also known as the Tanimoto distance.

***d* - Dice dissimilarity.**

Dice dissimilarity between two individuals x and y is determined as follows:

$$d(x, y) = \frac{b + c}{2a + b + c}. \quad (\text{A3})$$

The Dice coefficient of dissimilarity varies between 0 and 1.

Note: $1 - d(x, y)$ is also known as the **Nei and Li (1979) genetic similarity** measure and the **Sørensen** measure of similarity for composition of species in ecology.

2. Comparison of differentials (columns of binary data matrix).

***Cor* - Correlation between differentials.**

The measure of correlation between two differentials D_1 and D_2 is defined by the formula

$$\text{Cor}(D_1, D_2) = 1 - 2m(D_1, D_2), \quad (\text{A4})$$

$m(D_1, D_2)$ is value of the simple mismatch dissimilarity (A1) between binary vector-columns corresponding to differentials D_1 and D_2 (Kosman 2003b). This measure ranges from -1 to 1, $-1 \leq \text{Cor}(D_1, D_2) \leq 1$: $\text{Cor}(D_1, D_2) = 1$ if the two vector-columns are identical, and $\text{Cor}(D_1, D_2) = -1$ if binary representation of each individual is different on D_1 and D_2 (01 or 10). Differentiating characters are strongly correlated if $\text{Cor}(D_1, D_2)$ is relatively close to -1 or to 1, whereas there is no correlation between them if the values of $\text{Cor}(D_1, D_2)$ are close to zero.

***φ* - Association between differentiating characters.**

Coefficient of association ϕ between pairs of differentiating characters D_1 and D_2 is determined by formula 17.5 in Sokal and Rohlf (1995):

$$\phi(D_1, D_2) = \frac{\alpha\delta - \beta\gamma}{\sqrt{(\alpha + \beta)(\gamma + \delta)(\alpha + \gamma)(\beta + \delta)}}, \quad (\text{A5})$$

where α = number of individuals with shared positive responses to the both differentials D_1 and D_2 (11), β = number of individuals with positive response to D_1 , but negative response to D_2 (10), γ = number of individuals with negative response to D_1 , but positive response to D_2 (01), and δ = number of individuals with shared negative

responses to the both differentials D_1 and D_2 (00). Obviously, that $\alpha + \beta + \gamma + \delta = n$, where n is the total number of individuals tested. The ϕ -coefficient is related to Fisher's exact test (Sokal and Rohlf 1995).

3. Characterization of individual patterns.

Let us consider a dichotomous pattern of any individual I tested on a set of k differentiating factors with positive responses (1s) on n^+ of them. Then numbers of 1s and 0s in binary vector representing I equal n^+ and $n^- = k - n^+$, respectively.

C – individual Complexity.

Complexity of individual I is determined as the number of 1s (positive responses) in its binary representation:

$$C(I) = n^+. \quad (\text{A6})$$

RC – Relative individual Complexity.

Relative complexity of individual I is determined as its complexity normalized by the total number of differentiating factors:

$$RC(I) = \frac{n^+}{k}. \quad (\text{A7})$$

The relative complexity varies between 0 and 1, $0 \leq RC(I) \leq 1$.

U – individual Uniformity.

Uniformity of individual I is determined as follows (Kosman 2003b):

$$U(I) = 1 - 2 \cdot \min\{RC(I), 1 - RC(I)\}. \quad (\text{A8})$$

This measure ranges from 0 to 1, $0 \leq U(I) \leq 1$. If there are only 1s (or 0s) in the response pattern, then the uniformity is maximal and equals 1. The uniformity reaches its minimum value 0 if the numbers of positive (1s) and negative (0s) responses are equal.

U_I – sample Uniformity.

Measure of uniformity for a sample of n individuals I_j ($j=1,2,\dots,n$) from population P is defined as the average of the uniformity values (3) across individuals:

$$U_I(P) = \frac{1}{n} \cdot \sum_{j=1}^n U(I_j). \quad (\text{A9})$$

This average uniformity of individuals is called the sample/population uniformity. It ranges from 0 to 1, $0 \leq U_I(P) \leq 1$.

U_{ph} – average phenotype Uniformity.

Average uniformity of s phenotypes/genotypes ph_j ($j=1,2,\dots,s$) sampled from population P is determined as follows:

$$U_{ph}(P) = \frac{1}{s} \cdot \sum_{j=1}^s U(ph_j). \quad (A10)$$

It ranges from 0 to 1, $0 \leq U_{ph}(P) \leq 1$.

Designations for virulence data.

Virulence complexity, VC , relative virulence complexity, RVC , virulence uniformity of individual, VU , and average virulence uniformities of individuals (sample/population uniformity), VU_I , and phenotypes/genotypes, VU_{ph} , are determined according to formulae (A6), (A7), (A8), (A9) and (A10), respectively.

Designations and formulae for resistance data.

For binary resistance data 1 and 0 usually represent susceptibility and resistance, respectively, of an individual for any given pathogen.

Resistance complexity, CR , relative resistance complexity, RCR , resistance uniformity of individual, UR , and average resistance uniformities of individuals (sample/population uniformity), UR_I , and phenotypes/genotypes, UR_{ph} , are determined according to formulae (A11), (A12), (A13), (A14) and (A15), respectively:

$$CR(I) = n^-, \quad (A11)$$

$$RCR(I) = \frac{n^-}{k}, \quad (A12)$$

$$UR(I) = 1 - 2 \cdot \min\{RCR(I), 1 - RCR(I)\}, \quad (A13)$$

$$UR_I(P) = \frac{1}{n} \cdot \sum_{j=1}^n UR(I_j), \quad (A14)$$

$$UR_{ph}(P) = \frac{1}{s} \cdot \sum_{j=1}^s UR(ph_j). \quad (A15)$$

4. Diversity within and among populations.

Consider a sample from population P which consists of n individuals. We assume that all individuals are tested on k differentiating factors D_1, D_2, \dots, D_k , and represented by binary patterns of 1s and 0s for positive (e. g. virulence, susceptibility) and negative (e. g. avirulence, resistance) responses, respectively. We denote by q_i the frequency of appearance of 1 at the i -th differentiating factor D_i . For example, if the differentiating factors comprise a typical set of differential host lines used in virulence tests of plant pathogens, q_i would be the frequency of virulence in population P on the i -th differential

line. For dominant molecular markers or any diallelic loci q_i would be the frequency of dominant allele in the i -th locus in population P .

We denote T_r a group of individuals of type r . The frequency of individuals of type r in population P is denoted by p_r . Depending on the nature of the data, types of individuals may mean pathotypes, races, phenotypes, genotypes etc.

Diversity within population (*pattern-based methods*).

Let population P consists of n individuals x_1, x_2, \dots, x_n , and dissimilarity between the individuals is assessed using any measure ρ (e. g. simple mismatch m , Jaccard j , Dice d coefficients of dissimilarity).

ADW - Average Difference Within population.

The Average Difference Within population P with respect to dissimilarity ρ , $ADW_\rho(P)$, is determined as follows:

$$ADW_\rho(P) = \frac{1}{n^2} \cdot \sum_{i,j=1}^n \rho(x_i, x_j) . \quad (A16)$$

KW – Kosman diversity Within population.

The Kosman diversity $KW_\rho(P)$ within population P of n individuals is defined as follows:

$$KW_\rho(P) = \frac{1}{n} \cdot Ass_{\max}^\rho(P, P) . \quad (A17)$$

(Kosman 1996, Kosman and Leonard, 2007). For given population P and dissimilarity measure ρ , $KW_\rho(P) \geq ADW_\rho(P)$.

Diversity within population (*type-based methods*).

Frequency of individuals of a fixed type.

Let n_r be the number of individuals of type T_r from population P , which consists of n individuals x_1, x_2, \dots, x_n , and s is the total number of types of individuals observed in this population (e. g. number of pathotypes, races, phenotypes, denotypes). Then frequency of individuals of type T_r ($r=1,2,\dots,s$) in population P is determined as follows:

$$p_r = \frac{n_r}{n} , \quad (A18)$$

so that $p_1 + p_2 + \dots + p_s = 1$.

Theoretical standard error of p_r estimate (A18) is expressed by formula

$$SE(p_r) = \sqrt{\frac{p_r(1-p_r)}{n}} \quad (\text{A19})$$

for $r = 1, 2, \dots, s$.

G - Gleason richness within population.

The Gleason index of richness within population P is defined as follows:

$$G(P) = \frac{s-1}{\ln n} . \quad (\text{A20})$$

Si – Simpson diversity within population.

The Simpson index of diversity within population P (Simpson 1949) is defined as follows:

$$Si(P) = 1 - \sum_{r=1}^s p_r^2 , \quad (\text{A21})$$

Its values range between 0 and $1 - \frac{1}{s}$, $0 \leq Si(P) \leq 1 - \frac{1}{s}$.

St – Stoddart diversity within population.

Stoddart's index of diversity within population P (Stoddart 1983, Stoddart and Taylor 1988) is defined as follows:

$$St(P) = \frac{1}{\sum_{r=1}^s p_r^2} . \quad (\text{A22})$$

Its values range between 1 and s , $1 \leq St(P) \leq s$.

SH – Shannon diversity within population (Shannon-Wiener entropy).

The Shannon index of diversity within population P (Shannon-Wiener entropy, Shannon and Weaver 1949) is defined as follows:

$$SH(P) = - \sum_{r=1}^s p_r \ln p_r . \quad (\text{A23})$$

Values of Shannon's diversity index range between 0 and $\ln s$, $0 \leq Sh(P) \leq \ln s$.

E – Evenness of population.

Measure of population evenness is expressed by the ratio of the Shannon index (A23) to its maximum value $\ln s$, $2 \leq s \leq n$ (Sheldon 1969):

$$E(P) = \frac{SH(P)}{\ln s} = - \frac{1}{\ln s} \cdot \sum_{r=1}^s p_r \ln p_r . \quad (\text{A24})$$

This evenness parameter ranges between 0 and 1, $0 \leq E(P) \leq 1$.

Sh - Normalized Shannon diversity within population.

The Normalized Shannon index of diversity within population P is defined as follows:

$$Sh(P) = \frac{SH(P)}{\ln n} = -\frac{1}{\ln n} \cdot \sum_{r=1}^s p_r \ln p_r. \quad (A25)$$

This diversity index also ranges between 0 and 1, $0 \leq Sh(P) \leq 1$. Sh -index reflects both the evenness and richness of population because

$$Sh(P) = \frac{SH(P)}{\ln n} = \frac{SH(P)}{\ln s} \cdot \frac{\ln s}{\ln n} = E(P) \cdot \ln_n s,$$

where the second factor measures the population richness ($0 \leq \ln_n s \leq 1$, minimum and maximum value being obtained in a population in which all individuals are of the same type, $s = 1$, and different types, $s = n$, respectively).

Diversity within population (trait-based methods).**Frequency of positive response.**

Let population P consists of n individuals x_1, x_2, \dots, x_n tested on a set of k binary differentiating factor D_j , $j = 1, 2, \dots, k$, positive and negative responses being represented by 1 and 0, respectively. Frequencies of appearance of 1 at the i -th differentiating factor D_i (proportion of 1s in i -th column of binary matrix) for populations P is denoted by q_j , $j = 1, 2, \dots, k$. Theoretical standard error of q_j estimate is expressed by formula

$$SE(q_j) = \sqrt{\frac{q_j(1-q_j)}{n}} \quad (A26)$$

for $j = 1, 2, \dots, k$.

 H_S - Nei diversity within population.

The Nei's measure of the average gene diversity per locus H_S in population P (Nei 1973) is determined by the formula

$$H_S(P) = \frac{1}{k} \cdot \sum_{j=1}^k H_{Sj}(P) = \frac{1}{k} \cdot \sum_{j=1}^k [1 - q_j^2 - (1 - q_j)^2], \quad (A27)$$

where k is the total number of loci (differentiating factors), $H_{Sj}(P) = 1 - q_j^2 - (1 - q_j)^2$, and q_j and $1 - q_j$ are frequencies of the two alleles at the j -th diallelic locus (e. g. q_j =virulence frequency, $1 - q_j$ =resistance frequency). Nei's diversity H_S ranges between 0 and 0.5 for binary data, $0 \leq H_S(P) \leq \frac{1}{2}$.

It was proved (Kosman 2003a) that the Nei's measure of the average gene diversity per locus $H_S(P)$ and the index of average difference (A16) with respect to the simple mismatch coefficient are identical measures of diversity within population:

$$ADW_m(P) = H_S(P). \quad (A28)$$

K-diversity within population.

The K -index for measuring diversity within population is determined as follows:

$$K(P) = \frac{1}{k} \cdot \sum_{l=1}^k K_l(P) = \frac{1}{k} \cdot \sum_{l=1}^k \min\{2q_l, 2(1-q_l)\} \quad (A29)$$

(Manisterski *et al.* 2000), where q_l is frequency of positive response (appearance of 1) at l -th character in populations P , $l = 1, 2, \dots, k$, and $K(P) = \min\{2q_l, 2(1-q_l)\}$ is the diversity within population at differentiating character D_l . The K -index was designated H_{KDiv} in Manisterski *et al.* (2000).

Distance between populations (pattern-based methods).

Let two populations P_1 and P_2 consist of n and n^* individuals x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_{n^*} , respectively, and dissimilarity between the individuals is assessed using any measure ρ (e. g. simple mismatch m , Jaccard j , Dice d coefficients of dissimilarity).

DAD - Distance of Average Differences between populations.

The Distance of Average Differences between populations P_1 and P_2 with respect to dissimilarity ρ is defined as follows:

$$DAD_\rho(P_1, P_2) = ADB_\rho(P_1, P_2) - \frac{ADW_\rho(P_1) + ADW_\rho(P_2)}{2}, \quad (A30)$$

where

$$ADB_\rho(P_1, P_2) = \frac{1}{n \cdot n^*} \cdot \sum_{i,j=1}^{n, n^*} \rho(x_i, y_j),$$

$$ADW_\rho(P_1) = \frac{1}{n^2} \cdot \sum_{i,j=1}^n \rho(x_i, x_j) \quad \text{and} \quad ADW_\rho(P_2) = \frac{1}{(n^*)^2} \cdot \sum_{i,j=1}^{n^*} \rho(y_i, y_j).$$

KB – Kosman Distance between populations.

The Kosman distance $KB_\rho(P_1, P_2)$ between two populations P_1 and P_2 is defined for two samples of equal number individuals $n = n^*$ as follows:

$$KB_\rho(P_1, P_2) = \frac{1}{n} \cdot Ass_{\min}^\rho(P_1, P_2) \quad (A31)$$

(Kosman 1996, Kosman and Leonard, 2007). For given populations P_1 and P_2 and dissimilarity measure ρ , $KB_\rho(P_1, P_2) \leq ADB_\rho(P_1, P_2)$.

Distance between populations (*type-based methods*).

Let populations P_1 and P_2 consist of n and n^* individuals $x_{11}, x_{12}, \dots, x_{1n}$ and $x_{21}, x_{22}, \dots, x_{2n^*}$, respectively, n_{1r} and n_{2r} be the number of individuals of type T_r from populations P_1 and P_2 , respectively, and s is the total number of types of individuals observed in both populations. Then $p_{1r} = \frac{n_{1r}}{n}$ and $p_{2r} = \frac{n_{2r}}{n^*}$ ($r = 1, 2, \dots, s$) will denote the frequency of individuals of type T_r in populations P_1 and P_2 , respectively, so that $p_{i1} + p_{i2} + \dots + p_{is} = 1$, $i = 1, 2$.

R - Rogers distance between populations.

The Rogers distance between two populations P_1 and P_2 is determined as follows (Rogers 1972):

$$R(P_1, P_2) = \frac{1}{2} \cdot \sum_{r=1}^s |p_{1r} - p_{2r}|. \quad (\text{A32})$$

Values of the Rogers distance ranges between 0 and 1, $0 \leq R(P_1, P_2) \leq 1$.

Distance between populations (*trait-based methods*).

Let populations P_1 and P_2 consist of n and n^* individuals $x_{11}, x_{12}, \dots, x_{1n}$ and $x_{21}, x_{22}, \dots, x_{2n^*}$, respectively, all individuals are tested on a set of k binary differentiating factors D_j , $j = 1, 2, \dots, k$, and positive (e. g. virulence, susceptibility) and negative (e. g. avirulence, resistance) responses are represented by 1 and 0, respectively.

N_M - Nei minimum genetic distance between populations.

Nei's minimum genetic distance between two populations P_1 and P_2 (Nei 1972) is determined as follows:

$$N_M(P_1, P_2) = \frac{J_1 + J_2}{2} - J_{12}, \quad (\text{A33})$$

where

$$J_1 = \frac{1}{k} \cdot \sum_{i=1}^k [q_{1i}^2 + (1 - q_{1i})^2], \quad (\text{A34})$$

$$J_2 = \frac{1}{k} \cdot \sum_{i=1}^k [q_{2i}^2 + (1 - q_{2i})^2] \quad (\text{A35})$$

and

$$J_{12} = \frac{1}{k} \cdot \sum_{i=1}^k [q_{1i} \cdot q_{2i} + (1 - q_{1i}) \cdot (1 - q_{2i})] \quad (\text{A36})$$

for k dimorphic loci with frequencies q_{1i} and $1 - q_{1i}$, and q_{2i} and $1 - q_{2i}$ of the two alleles at the i -th locus in populations P_1 and P_2 , respectively. For binary data q_{1i} and q_{2i} are the frequencies of appearance of 1 at the i -th differentiating factor D_i for populations P_1 and P_2 , respectively.

It was proved (Kosman and Leonard 2007) that the Nei's minimum genetic distance (A33) can be represented as the distance of average differences between populations P_1 and P_2 (A30) with respect to the simple mismatch dissimilarity:

$$\begin{aligned} N_M(P_1, P_2) &= \frac{J_1 + J_2}{2} - J_{12} = \\ &= ADB_m(P_1, P_2) - \frac{ADW_m(P_1) + ADW_m(P_2)}{2} = DAD_m(P_1, P_2). \end{aligned} \quad (\text{A37})$$

N - Nei standard genetic distance between populations.

Nei's standard genetic distance between two populations P_1 and P_2 (Nei 1972, 1978) is defined as follows:

$$N(P_1, P_2) = \frac{\ln J_1 + \ln J_2}{2} - \ln J_{12} = -\ln \frac{J_{12}}{\sqrt{J_1 \cdot J_2}}, \quad (\text{A38})$$

where J_1 , J_2 and J_{12} are calculated according to equations (A34–A36). The Nei standard genetic distance (A38) can be expressed as the following function of the measures of average difference within and between populations P_1 and P_2 :

$$N(P_1, P_2) = -\ln \frac{J_{12}}{\sqrt{J_1 \cdot J_2}} = -\ln \frac{1 - ADB_m(P_1, P_2)}{\sqrt{[1 - ADW_m(P_1)] \cdot [1 - ADW_m(P_2)]}} \quad (\text{A39})$$

(Kosman and Leonard 2007).

NSE - Normalized Squared Euclidean distance between populations.

The normalized squared Euclidean distance between two populations P_1 and P_2 is determined on the basis of frequencies of positive responses of individuals on all differentiating factors D_s , $s = 1, 2, \dots, k$:

$$NSE(P_1, P_2) = \frac{1}{k} \cdot \sum_{s=1}^k (q_{s1} - q_{s2})^2. \quad (\text{A40})$$

One can prove that the distance of average differences between populations P_1 and P_2 (A30) with respect to the simple mismatch dissimilarity equals the Nei's minimum genetic distance (A33) and the normalized squared Euclidean distance (A40):

$$DAD_m(P_1, P_2) = N_M(P_1, P_2) = NSE(P_1, P_2), \quad (A41)$$

G_{ST} - Nei coefficient of differentiation among populations.

The Nei coefficient of differentiation G_{ST} among populations (Nei 1973) is equivalent to the definition for F_{ST} measure (Wright 1951), and being applied to two populations P_1 and P_2 is used for measuring distance between populations:

$$G_{ST}(P_1, P_2) = \frac{1}{k} \sum_{l=1}^k G_{STl}(P_1, P_2), \quad (A42)$$

where $G_{STl}(P_1, P_2) = \frac{H_{Tl} - \hat{H}_{Sl}}{H_{Tl}}$ is calculated as follows. For l -th differentiating factor

$$\hat{H}_{Sl} = \frac{H_{Sl}(P_1) + H_{Sl}(P_2)}{2},$$

where $H_{Sl}(P_1) = 1 - (q_{1l})^2 - (1 - q_{1l})^2$, $H_{Sl}(P_2) = 1 - (q_{2l})^2 - (1 - q_{2l})^2$ and $H_{Tl} = 1 - q_{\bullet l}^2 - (1 - q_{\bullet l})^2$ for $q_{\bullet l} = \frac{q_{1l} + q_{2l}}{2}$. Values of the G_{ST} index vary between 0 and 1, $0 \leq G_{ST}(P_1, P_2) \leq 1$. One can show that the Nei coefficient of differentiation may be represented in the form

$$G_{ST}(P_1, P_2) = \frac{1}{k} \cdot \sum_{l=1}^k \frac{(q_{1l} - q_{2l})^2}{(q_{1l} + q_{2l})[(1 - q_{1l}) + (1 - q_{2l})]}. \quad (A43)$$

MCD - Mean Character Difference between populations.

The mean character difference MCD (pp. 122-123 in Sneath, and Sokal 1973) is identical to the Rogers traits based distance (Rogers 1972) in the case of dichotomous characters (binary data). The mean character difference between two populations P_1 and P_2 tested at k binary differentiating characters D_1, D_2, \dots, D_k , is determined as follows:

$$MCD(P_1, P_2) = \frac{1}{k} \cdot \sum_{l=1}^k CD_l(P_1, P_2) = \frac{1}{k} \cdot \sum_{l=1}^k |q_{1l} - q_{2l}|, \quad (A44)$$

where q_{1l} and q_{2l} are frequencies of positive response (appearance of 1) at l -th character in populations P_1 and P_2 , respectively, $l = 1, 2, \dots, k$, and $CD_l(P_1, P_2) = |q_{1l} - q_{2l}|$ is the distance of character difference between the populations at differentiating character D_l .

5. Resampled data

Let N fictive samples of a fixed size (number of individuals) are generated by means of resampling - drawing randomly with replacement from an original set of

individuals. If Par_i ($i = 1, 2, \dots, N$) is any parameter calculated for each fictive sample, then the mean estimator of the population parameter Par and its standard error SE are obtained as follows:

$$Par = \frac{1}{N} \sum_{i=1}^N Par_i, \quad (A45)$$

$$SE(Par) = \sqrt{\frac{\sum_{i=1}^N (Par - Par_i)^2}{N(N-1)}}. \quad (A46)$$

§8. References

1. Gilmour, J. 1973. Octal notation for designating physiologic races of plant pathogens. *Nature* 242: 620.
2. Habgood, R.M. 1970. Designation of physiological races of plant pathogens. *Nature* 227: 1268 – 1269.
3. Kosman, E. 1996. Difference and diversity of plant pathogen populations: A new approach for measuring. *Phytopathology* 86:1152-1155.
4. Kosman, E. 2003a. Nei's gene diversity and the index of average differences are identical measures of diversity within populations. *Plant Pathology* 52: 533-535.
5. Kosman, E. 2003b. Measure of multilocus correlation as a new parameter for study of plant pathogen populations. *Phytopathology* 93: 1464-1470.
6. Kosman E., Leonard K. J. 2007. Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction. *New Phytologist* 174: 683-696.
7. Manisterski, J., Eyal, Z., Ben-Yehuda, P., and Kosman, E. 2000. Comparative analysis of indices in the study of virulence diversity between and within populations of *Puccinia recondita* f. sp. *tritici* in Israel. *Phytopathology* 90: 601-607.
8. Nei, M. 1972. Genetic distance between populations. *Amer. Naturalist* 106:283-292.
9. Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70:3321-3323.
10. Nei, M. 1978. Estimation of average heterozygosities and genetic distance from a small number of individuals. *Genetics* 89: 583-590.
11. Nei, M., and Li, W. H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleasis. *Proceedings of the National Academy of Sciences of the USA* 76:5269-5273.
12. Pielou, E. C. 1974. *Population and Community Ecology. Principles and Methods*. Gordon and Breach, New York.
13. Roelfs, A. P., and D. V. McVey. 1973. Races of *Puccinia graminis* f. sp. *tritici* in the USA during 1972. *Plant Dis. Rep.* 57: 880-884.
14. Rogers, J. S. 1972. Measures of genetic similarity and genetic distance. Pages 145-153 in: *Studies in Genetics VII*. University of Texas Publication 7213, Austin.

15. Shannon, C. E., and Weaver, W. 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana.
16. Sheldon, A. L. 1969. Equitability indices: Dependence on the species count. *Ecology* 50:466-467.
17. Simpson, E. H. 1949. Measurement of diversity. *Nature* 163:688.
18. Sneath, P. A., & Soal, R. R. 1973. Numerical Taxonomy. W. H. Freeman Co., San Francisco.
19. Sokal, R. R., and Rohlf, F. J. 1995. Biometry. W. H. Freeman, New York.
20. Stoddart, J. A. 1983. A genotypic diversity measure. *J. Hered.* 74: 489-490.
21. Stoddart, J. A., and Taylor, J. F. 1988. Genotypic diversity: estimation and prediction in samples. *Genetics* 118:705-711.
22. Wright, S. 1951. The genetic structure of populations. *Ann.of Eugenics* 15:323-354.